Towards the Design of an oligoTEA Endosomal Escape Agent: Investigations in Hydrophobicity and pK_a

A Thesis
Presented to the Faculty of the Graduate School
Of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Masters of Science

by
Maximilian Minhthong Nguyen
December 2018

Acknowledgements

I would like to thank members of the Alabi lab for their help in conducting experiments, data analysis, and general support. In particular, I would like to thank Joseph Brown, Joshua Walker, and Mintu Porel for their assistance in using the LC-MS machine. I would like to thank Susan Daniel and Benjamin Widom for their helpful discussions. Special thanks to Christopher Alabi for helping me define a thesis project and the support and guidance to see it through. This work was supported by startup research funds from Cornell University and the Nancy and Peter Meinig Investigator Fellowship.

Biological Sketch

Hailing from the state of Georgia, Maximilian Nguyen comes to Cornell and Christopher Alabi's lab with undergraduate training in Chemical and Biomolecular Engineering from Georgia Tech. He has strong interests in probing biological systems from a basic science perspective. Given his background as an engineer, his interest in joining the Alabi group stems from also having a strong interest in translating fundamental biology to affect human health. He works at the interface of chemistry, biology, physics, and engineering, making problems in drug delivery a natural playground for exploration.

Table of Contents

Abstract

The oligothioetheramide (oligoTEA) family is a novel set of synthetic macromolecules with the advantageous characteristic of sequence-specificity. Due to their abiotic and highly modular nature, oligoTEAs are currently being explored in a variety of biological applications, ranging from use as antibacterial and antiviral agents, to use as heteromultifunctional cross-linkers and cell-penetrating agents.

The problem of endosomal escape in intracellular drug delivery represents an arena where this family of macromolecules may find potential use. Here, I report initial investigations into the physiochemical properties of oligoTEAs that would be relevant in tackling such a problem. Together, these methods and results can form the basis for future rational design of an oligoTEA endosomal escape agent.

pK_a and hydrophobicity are two key parameters routinely used in assessing the viability of new drug delivery candidates. Due to the key role of protonation in activating the function of a hypothetical endosomal escape agent, a study was performed to characterize the impact of certain noncovalent intramolecular interactions on side-chain protonation. From partitioning data, bulk microscopic measurements of pK_a and hydrophobicity were determined for a library consisting of oligoTEAs of differing length and backbone composition.

Polyprotic oligoTEAs may differ in their ionization potential and endosomal escape potency compared to their monoprotic counterparts if there are significant side-chain interactions. The presence or absence of nanoscale intramolecular interactions between oligoTEA sidechains is inferred using equilibrium statistical mechanical models on the partition measurements. In doing so, I highlight the difficulty in using molecular topology to predict the ability of a side-chain to influence the protonation of neighboring side-chains.

Rationalizing and interpreting these physiochemical properties of oligoTEAs is taken from a bottom-up perspective. Correlations between the parameter values of individual components and the parameter values of composite structures are highlighted. The data suggests that interactions between the side-chain and backbone interactions in the class of oligoTEA macromolecules explored in this work are minimal.

Introduction

Progress of RNAi therapeutics

It has been nearly 2 decades since the endogenous gene silencing mechanism known as RNA interference (RNAi)[1] was demonstrated to occur in mammalian cells[2]. Soon after, RNAi was shown to have therapeutic potential in mice[3] and in humans[4]. Since then, considerable effort has been expended to translate this powerful genetic tool into targeted therapeutics. Despite several setbacks, research and development in this area seems to be gaining steam. As of 2017, 16 companies were pursuing RNAi drugs across 28 clinical trials[1,2]. While there has been an estimated billions spent in developing RNAi therapeutic technologies[6], to date, only one RNAi drug has gained FDA approval[7,8]. While the first siRNA therapy, patisiran by Alnylam, which was approved August 2018, is certainly a milestone breakthrough in the area of RNA therapeutics, a number of key challenges remain to be addressed by the community[9–12].

Lipid nanoparticle-encapsulated and GalNAc-conjugated siRNAs are the current industry standard for RNAi drug delivery[13]. A common drawback is that both platforms almost exclusively target liver cells. Looking to deliver RNAi drugs to other organs necessitates exploration of next-generation strategies. A key problem is that there is no known organ-specific ligand-receptor pair that both expresses at the level (in excess of 500,000 surface copies) and has the low cycling time (15-20 minutes) of the GalNAc-ASGPR pair on hepatocytes[14–17]. While it may be easy for hepatocyte targets to reach the estimated ~2-5,000 siRNA copies needed to elicit a RNAi response[18], the cytosolic RNA copy number becomes a major issue when moving to targets beyond liver cells. Thus, it becomes a problem of either (1) finding better ligand-receptor pairs that are both abundant and rapidly cycling or (2) improving the efficiency of siRNA uptake and escape from the endocytic pathway. This work is broadly motivated by the engineering challenges presented by the latter.

Barriers to siRNA delivery: endosomal escape

Uptake across the cell membrane of exogenous double-stranded RNA (dsRNA), the key drug payload in RNAi therapeutics, is known to be mediated via receptor-mediated endocytosis[19,20]. Several leaders in the field make the case that controlled release of dsRNA from the endocytic pathway (termed endosomal escape) remains the greatest obstacle to the next generation of RNAi drugs[21–25]. Delivery vectors should be stable, low-toxicity agents that can facilitate efficient, reliable

escape of dsRNA cargo out of the endocytic pathway before the cargo becomes subject to lysosomal degradation. An apt analogy for the problem is the use of the Trojan Horse to deliver a payload of Greek soldiers to invade the city of Troy. Here, the payload is dsRNA that needs to make it into the cell without being detected, and the Horse itself is any delivery system that will allow unimpeded entry into the cell's interior, with the additional caveat that the Horse must enter the cell through a specific entryway (the endocytic pathway). Once inside the cell, the problem of endosomal escape means there needs to be a mechanism to release the payload from the Horse or else the cell will just destroy the Horse after a certain amount of time.

It is interesting to note that Nature has developed solutions to this problem; many bacteria and viruses have evolved mechanisms to facilitate endosomal escape. One famous example is the hemagglutainin protein found in influenza viruses[26]. Reports of their possible mechanisms include strategies such as pore formation, the proton sponge effect, and membrane fusion[22,27]. But rather than simply extracting the specific compounds bacteria and viruses use or making similar derivatives, a more fruitful and scalable enterprise would be to understand the principles behind those mechanisms and use that inspiration to design a broader class of endosomal escape agents. For instance, the use of pH in some of these mechanisms as a control switch is the source of inspiration for the synthetic escape agent pursued in this work.

Currently there are three broad strategies to facilitate endosomal escape: use of membrane-destabilizing lipids, use of membrane-destabilizing peptides/polymers, and increasing endosomal accumulation[9]. The use of peptides and polymers as membrane-destabilizing agents is attractive due to the vast number of possible chemical modifications that can be made[28–30]. Synthetic polymer-based agents typically have the additional advantage of being less susceptible to degradation. Careful choice of chemical modification can also prevent invoking an immune response, as demonstrated in a number of synthetic systems[31,32]. With the large selection of potential chemistries, the goal becomes finding a chemistry toolbox that will allow for precise engineering of an endosomal escape function.

OligoTEA chemistry: a modular platform for engineering sequence-specific function

Oligothioetheramides (oligoTEAs) are a novel family of sequence-defined compounds incorporating the efficiencies of click chemistry[33]. Using a unique allyl acrylamide monomer with two orthogonal reactive sites, sequence-specific oligomer growth using alternating thiol-Michael and

thiol-ene click additions can be rapidly achieved without the need for protection/deprotection strategies. This robust platform allows for a rich diversity of functional groups in both the allyl acrylamide monomer side chain and in the dithiol comonomer. Consequently, the combinatorial nature of their diversity offers potential for engineering sequence-specific function into biotechnology applications. Due to their abiotic and highly modular nature, oligoTEAs are currently being explored in a variety of biological settings, ranging from use as antibacterial[34,35] and antiviral agents, to use as heteromultifunctional cross-linkers[36] and cell-penetrating agents.

As the previous sections suggest, one other arena where this family of compounds may find potential use is the problem of endosomal escape in intracellular drug delivery. Broadly speaking, the pharmaceutical industry is approximately a $1 trillion global market, which would easily explain why there is an enormous literature of innovation on the issue of drug delivery (IMS Institute for Healthcare Information. Global Outlook for Medicines Through 2018, 2014). In the domain of intracellular delivery, such as in the systemic delivery of RNA, while the amount of effort has been extensive, few satisfying solutions have made it through the clinical process[37]. OligoTEAs may be a versatile ingredient in addressing this problem as they may be used in several ways: (1) complexing the oligoTEA to a drug to make a formulation, (2) using the oligoTEA in a liposome system, or (3) conjugating the oligoTEA directly to the drug.

For promising young technologies such as the CRISPR-Cas9 system, the question of delivery is still in an infantile stage of exploration[38]. While the focus of this work was done with applications for siRNA delivery in mind, the problem of endosomal escape is a generic one for intracellular delivery. I present here initial investigations in using oligoTEAs as a platform for tackling endosomal escape. Together, these methods and results can be taken as part of a basis for future rational design of an oligoTEA endosomal escape agent.

Part 1. Designing an oligoTEA library to test noncovalent intramolecular interactions

Key Design Parameters: pK_a and hydrophobicity

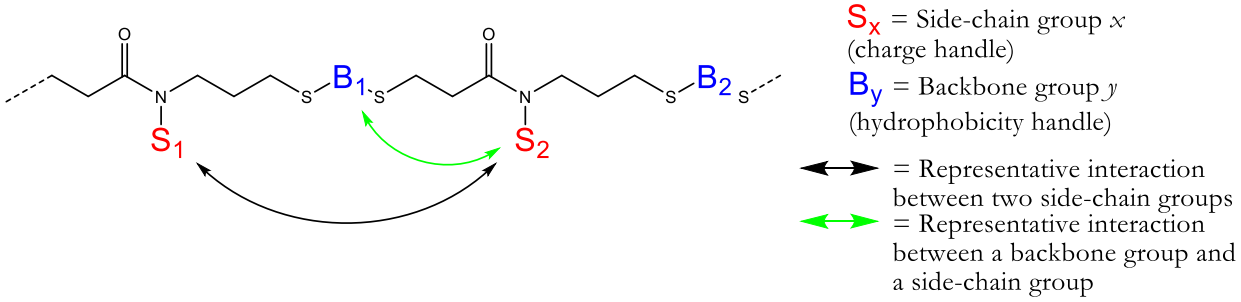
The acidity of endosomal compartments has been observed to be highly dynamic, decreasing from a physiological pH of ~7.4 to a pH of ~5 on the order of minutes after endosome formation[39]. Some delivery strategies seek to exploit this endosomal maturation to mask the system's presence in the endocytic pathway until the endosome reaches a pH threshold. Subsequent acidification of the endosomal compartment triggers a change (typically conformational) in the system, activating its endosomal escape function. The acid dissociation constant, known commonly as the pK_a, reflects the lability of a compound to be protonated, making it a natural parameter to tune to respond to the endosome environment. Studies show the pK_a of a membrane-destabilizing compound can be a critical parameter in delivery efficacy[40–42]. This suggests kinetic and spatial control of the protonation reaction in this pH window as a potentially important design consideration for endosomal escape agents.

The hydrophobicity is another key parameter to consider in designing a potential escape agent. Lipophilicity, a term often associated with hydrophobicity, has long been known to be implicated in drug properties such as adsorption, distribution, metabolism, elimination, and toxicology[43]. From a physical standpoint, escaping the endosome necessitates an intimate interaction with a lipid membrane[44]. The hydrophobicity of an agent can serve as a potential proxy for predicting that membrane interaction. From a synthesis standpoint, the challenge lies in understanding how the overall hydrophobicity emerges from the different chemical components of the agent.

If one imagines a pK_a and hydrophobicity parameter space that contains every possible oligoTEA sequence, one would of course want to pick parameters that ultimately lead to optimal escape of the payload. The problem is that both (1) the values of these optimal parameters are unknown and (2) it is not immediately clear how parameter values change as you make modifications to the oligoTEA components.

Exploring oligoTEA noncovalent intramolecular interactions

Quantifying the effect of noncovalent intramolecular interactions between components of an oligoTEA should provide an initial thrust for more rational movement in this design space. The interactions, especially if charged groups are involved, will help dictate the oligoTEA's conformation and dynamics in solution. This is critical because how the ionized side chains are presented to other molecules in solution will dictate the nature of that interaction. While the mechanism of the membrane-destabilizing interaction believed to occur during endosomal escape is not well characterized[30], the conformation of the oligoTEA and the charges it presents will affect the intermolecular interactions with the endosomal membrane[45]. Thus, understanding noncovalent interactions would also aide in the prediction of intermolecular interactions between the escape agent and endosomal membrane. Due to the key role of protonation in activating the escape function, two logical subsets of intramolecular interactions to study are the effects of neighboring side-chains and backbone groups on the protonation of an oligoTEA side-chain.



Interactions between two side-chains groups

Polyprotic compounds may differ in their endosomal escape potency compared to their monoprotic counterparts if there are significant side-chain interactions. Here, I begin to address how molecular topology might influence the protonation of neighboring side-chains. A cation can be generated by choosing a side-chain group (such as a tertiary amine) that can be protonated at pH's resembling an endosome. The sequence of an oligoTEA can thought of as a molecular topology that gives a connectivity distance between side-chain groups. By using different backbone comonomers, one can purposely modulate the topological distance between neighboring side-chains. Thus, one question we would like to answer is what effect does the molecular connectivity have on electrostatic interactions, and how does that in turn affect the molecule's pK_a?

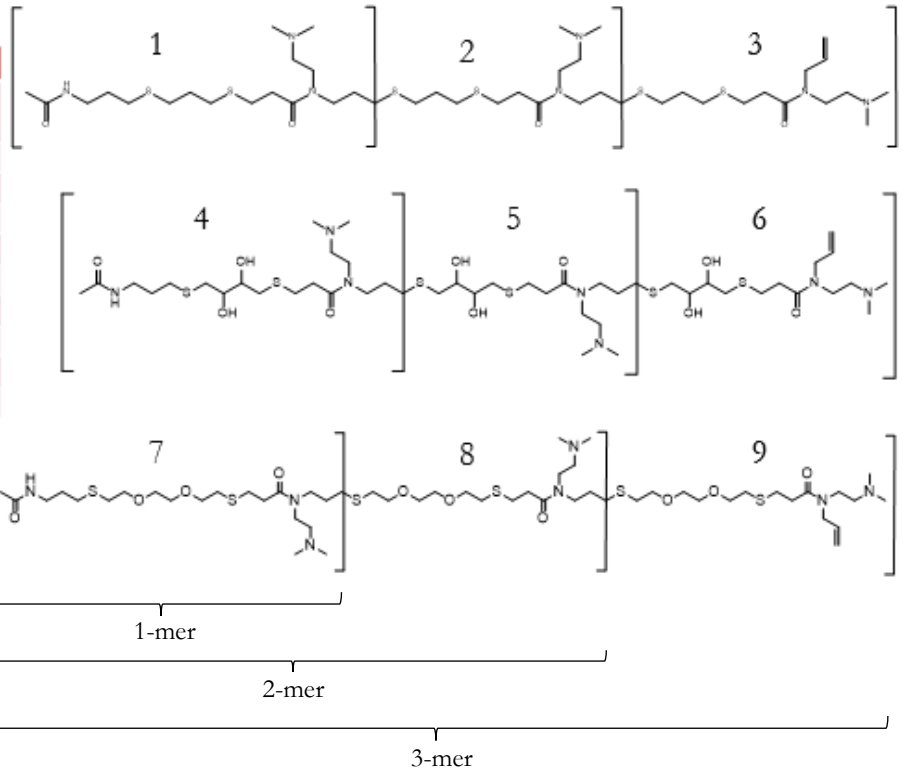
Interactions between the backbone and a side-chain group

How is the ionization of an ionized side-chain group affected by the hydrophobicity of the backbone group? A recent study suggests charge and hydrophobicity are closely linked[46], which suggests that the composition and hydrophobicity of the backbone should affect the electrostatic environment surrounding the side-chain moiety. Depending on how the backbone perturbs the local environment, this could cause slightly different molecular conformations, which in principle could affect the likelihood for the side-chain group to be protonated. If true, by varying the choice of backbone group, I would have an additional handle to be able to move in the oligoTEA pK_a and hydrophobicity space.

Through varying the mer length and molecular topology, one can presumably modulate the interaction between side-chains. By choice of backbone group, one influences the interaction between a side-chain group and the oligoTEA backbone. Therefore, I chose a library consisting of 1-, 2-, and 3- mers to observe interactions between side-chains. The 1- mer was used as a negative control since there is no side-chain interaction possible in that compound. There were 3 different dithiol backbone comonomers, each of a different hydrophobicity: a hydrophobic comonomer (propanedithiol), a hydrophilic comonomer (D/L-dithiothreitol), and an amphiphatic comonomer (2,2'-(Ethylenedioxy)diethanethiol). The resulting 3x3 library can be visualized as follows.

Dithiol comonomer	Relative Hydrophobicity
Propanedithiol (PDT)	Hydrophobic
D/L-Dithiothreitol (DTT)	Hydrophilic
2,2′ -(Ethylenedioxy)diethanethiol (PEG2)	Amphiphatic

Original Library		
Library Number	Oligomer Length	Dithiol Comonomer
1	1	PDT
2	2	PDT
3	3	PDT
4	1	DTT
5	2	DTT
6	3	DTT
7	1	PEG2
8	2	PEG2
9	3	PEG2



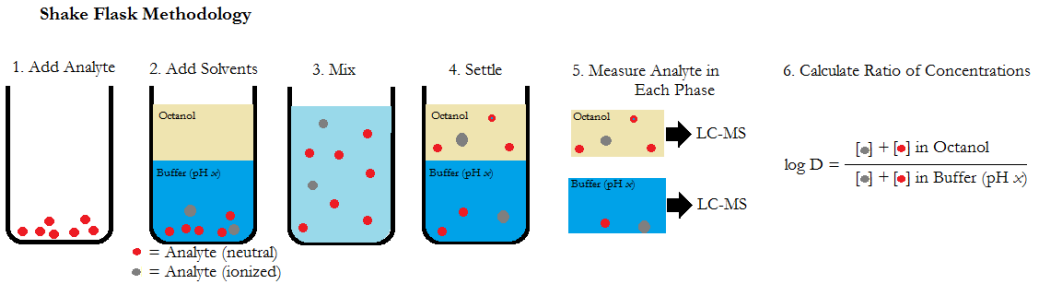
The oligoTEA library was synthesized using protocols adopted from Porel et. al[33], purified using RP-HPLC, and characterized with both LC-MS and NMR. The protocols are outlined in the Materials and Methods section and the relevant spectra are given in the Compound Verification section (Supplementary Info).

Part 2. Parameter analysis at bulk microscopic resolution from partition measurements

Shake flask method for log D measurements

For measuring the hydrophobicity and pK_a of a compound, there is a convenient experiment to get at both parameters simultaneously. Termed the shake flask method, this experiment measures the equilibrium partitioning of a compound between two immiscible phases (ex. octanol and water). Due to synthesis scale constraints, the measurements require high sensitivity. Using a highly sensitive and high-resolution technique such as LC-MS, one can accurately quantify concentrations as low as ~100 ng/mL in each phase. The relative distribution of compound between the two phases is termed the distribution coefficient (abbreviated D and generally reported in the literature in log form, hence log D measurements). A schematic of the methodology is given below.

By repeating the experiments in aqueous solutions of different pH, one can correlate the molecule's



ionization profile with its hydrophobicity. From that profile, one can extract out parameters such as pK_a and different aspects of hydrophobicity. The measurement is considered a bulk measurement because only the total concentration of all ionization species in a phase is measured (rather than the concentration of each individual ionization species if there are multiple ionization states, ex. the multi-mers may have multiple ionization states).

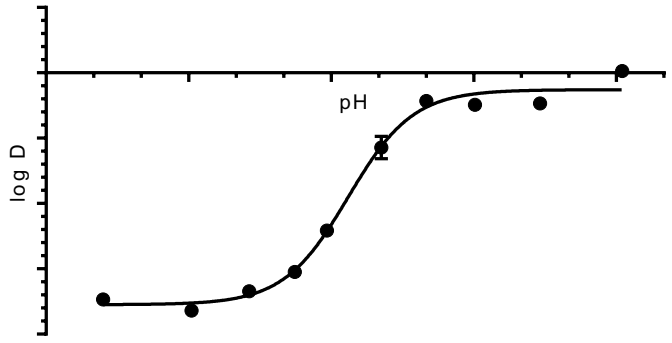
Since the D parameter is an equilibrium thermodynamic parameter, it can also be thought of as the equilibrium constant associated with the free energy transfer of a compound into the octanol phase from the aqueous phase.

$$\text{Compound}\,(aqueous) \rightleftarrows \text{Compound}\,(octanol) \qquad \Delta G° = RT\ln D$$

Only recently has the log D parameter been used as a useful measure of hydrophobicity for peptide-mimetics. Bolt et. al showed that hydrophobicity quantified through log D can be more reflective of the proper folded state of a compound than hydrophobicity quantified through HPLC-retention time[47]. Quantifying hydrophobicity from reverse-phase HPLC retention times offers different information because those runs are generally done at highly acidic conditions (pH <2), which would

not properly reflect conditions seen in a biological setting. Whereas in a partition experiment, the choice of the contents of each phase can be arbitrarily chosen. These experiments are the first log D measurements on the oligoTEA class of compounds.

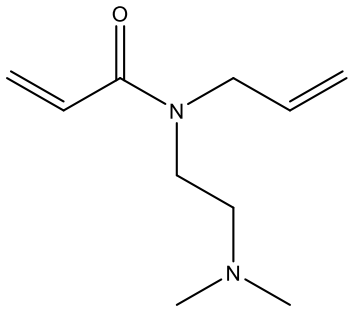
Shown below is a log D vs. pH plot characteristic of the data obtained from a shake flask experiment. The curves on these plots can be naturally fit with a sigmoidal function with 4 independent parameters: pK_a, Hill slope, log D_{max}, and log D_{min}. The pK_a occurs at the inflection point in the sigmoid, which is physically interpreted as the concentration of protons in solution (i.e. the pH) where the number of neutral species is equal to the number of ionized species. The Hill slope is the linear slope of the sigmoid at the inflection point; remembering that log D can be defined in terms of the free energy difference to transfer the compound between the two phases, the Hill slope can be thought as physically corresponding to the greatest possible free energy difference for transferring the compound between the two phases across all pH conditions. log D_{max}, and log D_{min} correspond to the lower and upper plateau values of the sigmoid respectively. Physically, log D_{min} is the distribution of the compound into the two phases, where all the molecules of the compound are in the maximally ionized state. While conversely, log D_{max} is where all the molecules of the compound are in the neutral state. A fifth parameter, the span of log D, can be defined as the difference between the log D_{max} and log D_{min}.



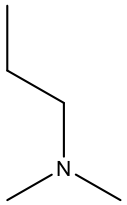
Characterizing the side chain: a reference point

In all members of the library, the only location where protonation can occur is on the N,N-dimethylethylamine subunit, referred to henceforth as the tertiary amine, of the allyl acrylamide monomer. If one considers an entire oligomer as a string of individual allyl acrylamide and comonomer components, then the global features of the entire oligomer should emerge from

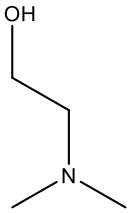
knowledge of the individual components (i.e. the sequence-to-structure-to-function paradigm). I characterized the pK_a and hydrophobicity of the monomer and its side chain to serve as a baseline reference. The ideal case would have been to characterize N,N-dimethylethylamine for the side chain, but its high volatility (a boiling point of 37 °C) made it difficult to handle experimentally. Instead, an alcohol variant (N,N-dimethylethanolamine) was used as a proxy.



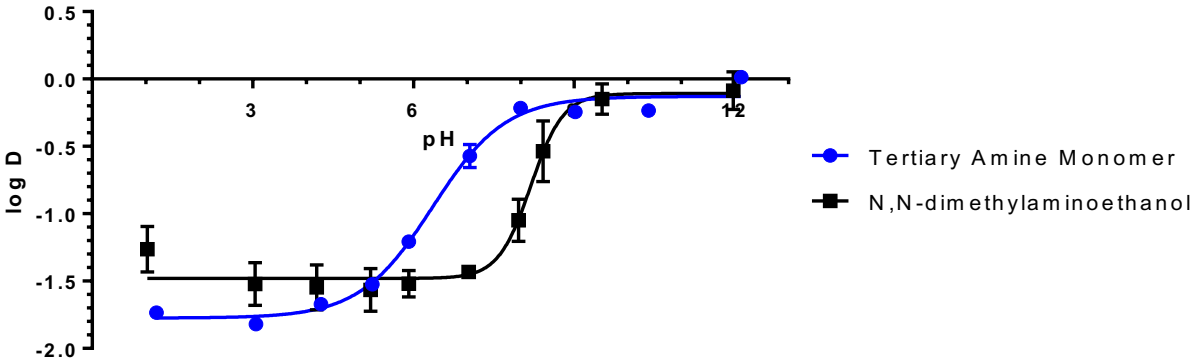
Tertiary amine monomer



N,N-dimethylethylamine



N,N-dimethylethanolamine



Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
Tertiary amine monomer	6.35±0.09	0.64±0.07	-0.13±0.03	-1.77±0.04	1.65±0.05
N,N-dimethylaminoethanol	8.19±0.09	1.43±0.47	-0.11±0.08	-1.48±0.04	1.37±0.09

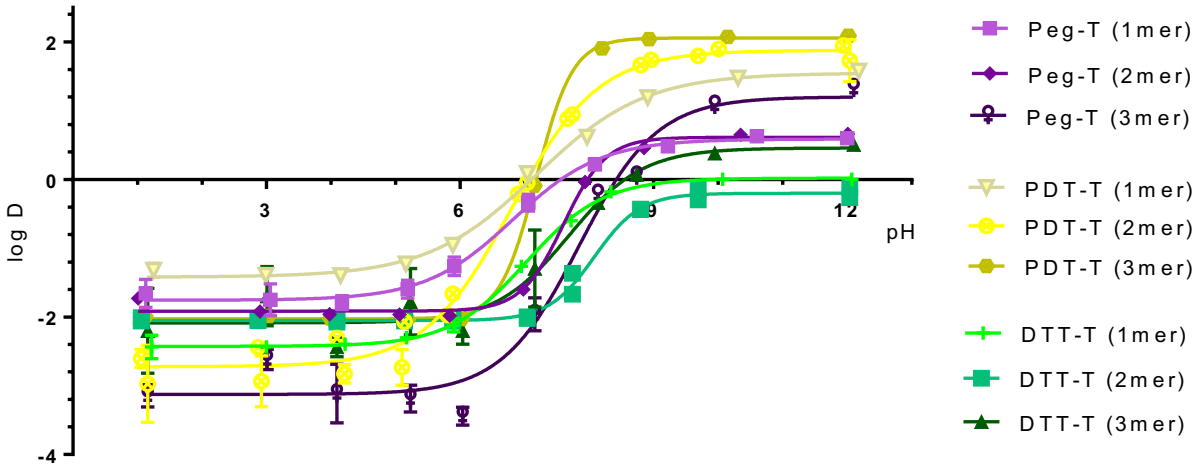
In contrast with aliphatic tertiary amine compounds[48] which typically have a pK_a of ~10, it is interesting to note that the pK_a of the tertiary amine monomer (T) is around 6.35. Fortunately for applications in endosomal escape, this pK_a happens to fall within the pH maturation window of the endosome, pH 5 to pH 7.4.

Both the subunit and the full monomer have nearly the same log D_{max} (which is also referred to as the log P) when they are both uncharged. The drastically different pK_a's of 6.4 versus 8.2

suggest that the local environment surrounding the tertiary amine may have a significant impact on its ability to be protonated. Evidence that local hydrophobicity can effect pK_a has previously been seen in amino lipid studies[41]. Zhang et. al observed that a more hydrophobic environment generally leads to lower pK_a values than expected from an isolated moiety. Recent force microscopy studies have also shown that increasing charge can increase local hydrophobicity[46,49].

Characterizing the library

With knowledge of the pK_a of the monomer as a reference, I then looked at the pK_a and hydrophobicities of a library of oligomers. The following shows the log D curves for all 9 members of the library as a function of pH. An aggregate plot of all members is included, and plots with subsets of the library are shown for clarity. I make note of qualitative trends and correlations seen in the data but leave aside speculation to their explanations for now.

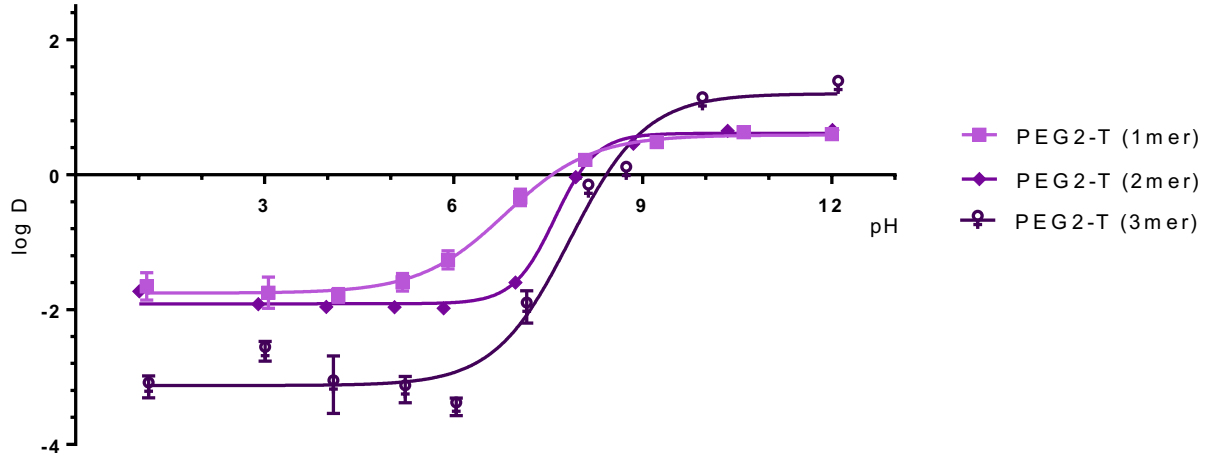


Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
PEG2-T (1mer)	6.82±0.08	0.64±0.06	0.59±0.04	-1.76±0.04	2.34±0.06
PEG2-T (2mer)	7.61±0.04	1.33±0.13	0.61±0.04	-1.92±0.03	2.53±0.05
PEG2-T (3mer)	7.85±0.13	0.72±0.13	1.20±0.19	-3.13±0.12	4.33±0.24
PDT-T (1mer)	7.18±0.07	0.52±0.04	1.55±0.05	-1.42±0.04	2.97±0.07
PDT-T (2mer)	6.80±0.06	0.64±0.05	1.88±0.08	-2.72±0.07	4.60±0.11
PDT-T (3mer)	7.18±0.01	1.55±0.12	2.06±0.02	-2.03±0.01	4.08±0.02
DTT-T (1mer)	7.02±0.05	0.73±0.05	0.02±0.03	-2.43±0.03	2.45±0.05

DTT-T (2mer)	8.08±0.04	1.22±0.09	-0.20±0.03	-2.05±0.01	1.85±0.03
DTT-T (3mer)	7.69±0.23	0.76±0.25	0.46±0.19	-2.09±0.13	2.54±0.24

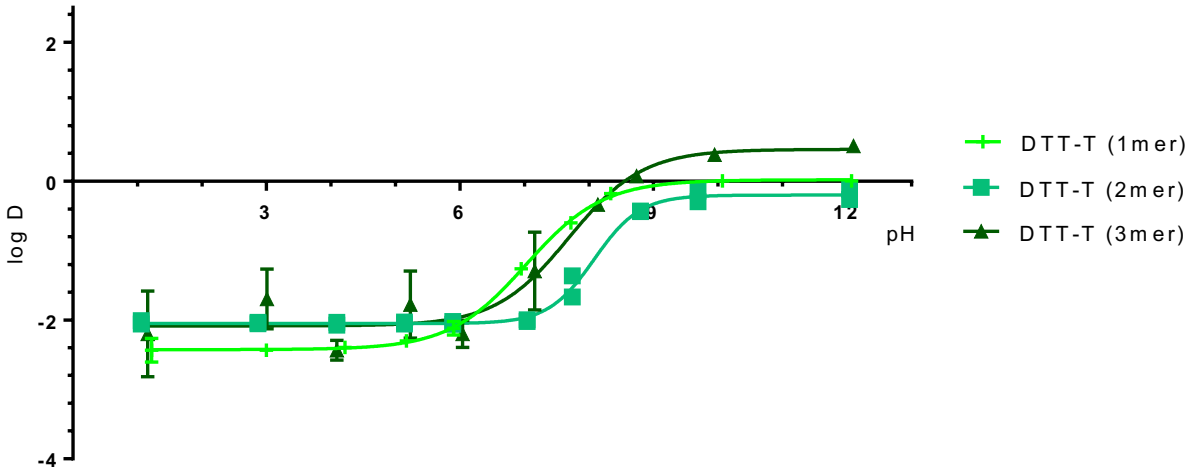
The pK_a's fall within the range of 6.8-8.1. Thus, in all cases, the apparent pK_a of the compound was greater than the pK_a of the individual tertiary amine monomer. Even if the compound has multiple charges, it is important to note that these measurements give a $pK_{a,\,average}$. Even though each side chain component is identical, in principle the compound might have a unique pK_a for each multiple of charge. This single effective $pK_{a,\,average}$ is due to the nature of the measurements. While there may be multiple species of different ionization states during the measurement at each pH condition, only an ensemble measurement over the partitioning of all species can be made.

By mer length



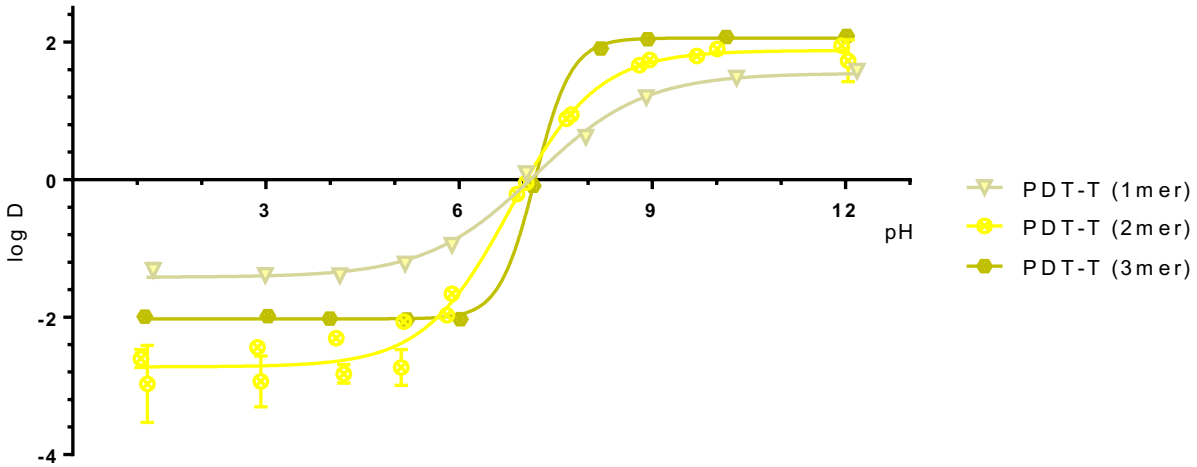
Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
PEG2-T (1mer)	6.82±0.08	0.64±0.06	0.59±0.04	-1.76±0.04	2.34±0.06
PEG2-T (2mer)	7.61±0.04	1.33±0.13	0.61±0.04	-1.92±0.03	2.53±0.05
PEG2-T (3mer)	7.85±0.13	0.72±0.13	1.20±0.19	-3.13±0.12	4.33±0.24

There are a couple straightforward correlations with the PEG2 family of compounds. The pK_a goes up, log P increases, log D_{min} decreases, and the span of log D increases with increasing mer length. In particular, the increase in length from 2- to 3- mer greatly enlarges the hydrophobicity profile.



Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
DTT-T (1mer)	7.02±0.05	0.73±0.05	0.02±0.03	-2.43±0.03	2.45±0.05
DTT-T (2mer)	8.08±0.04	1.22±0.09	-0.20±0.03	-2.05±0.01	1.85±0.03
DTT-T (3mer)	7.69±0.23	0.76±0.25	0.46±0.19	-2.09±0.13	2.54±0.24

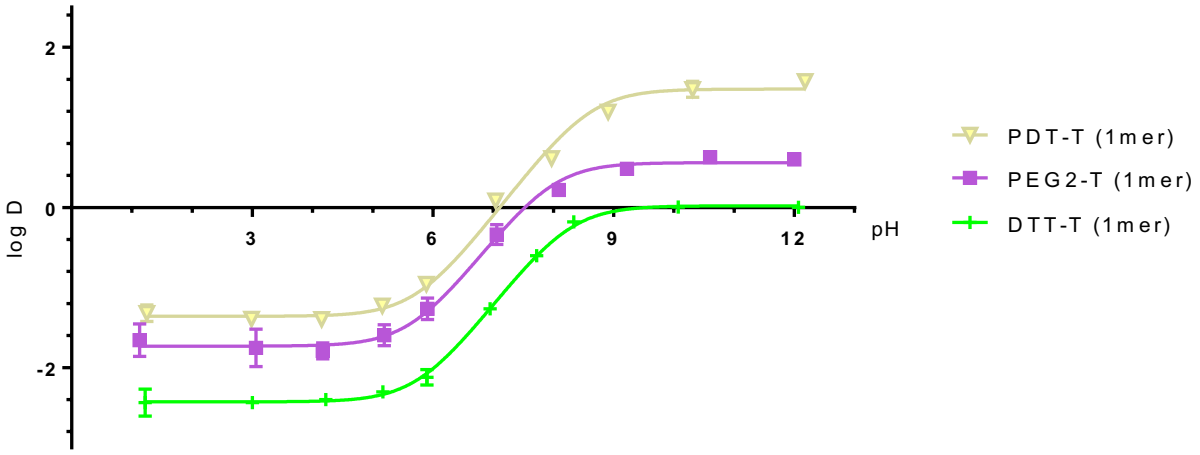
The DTT family seems to have no glaring correlations. Qualitatively, the three profiles appear to be fairly superimposable.



Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
PDT-T (1mer)	7.18±0.07	0.52±0.04	1.55±0.05	-1.42±0.04	2.97±0.07
PDT-T (2mer)	6.80±0.06	0.64±0.05	1.88±0.08	-2.72±0.07	4.60±0.11
PDT-T (3mer)	7.18±0.01	1.55±0.12	2.06±0.02	-2.03±0.01	4.08±0.02

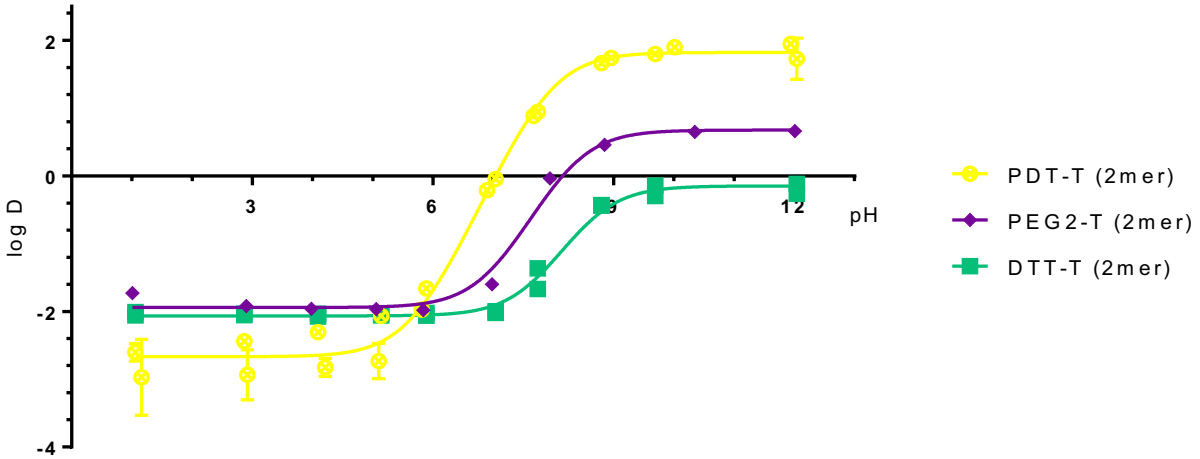
For the PDT family, it seems the span of hydrophobicity of multiple mers is substantially increased from the single mer. There's an increase in Hill slope and log P with mer length, but the pK_a of all three are fairly close.

By dithiol backbone comonomer



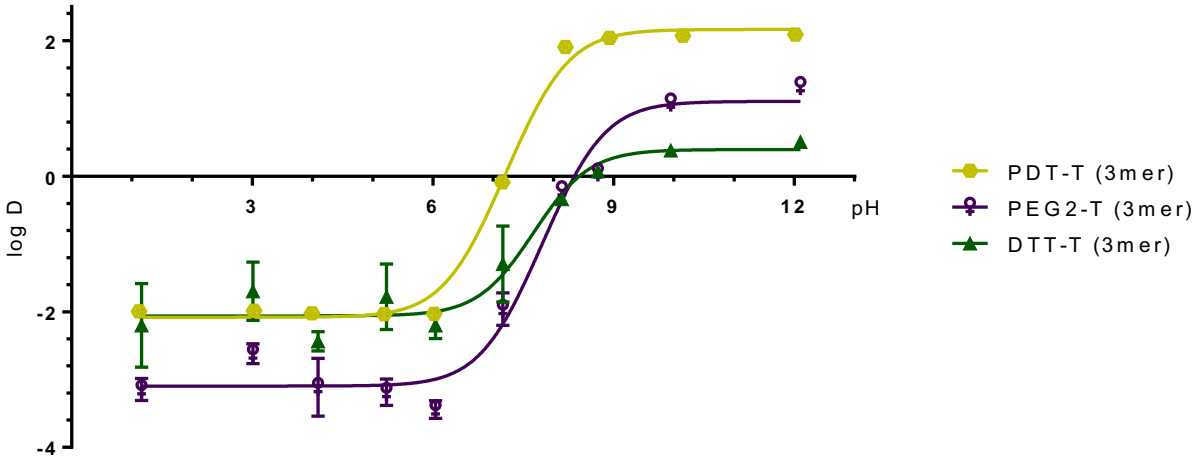
Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
PDT-T (1mer)	7.18±0.07	0.52±0.04	1.55±0.05	-1.42±0.04	2.97±0.07
PEG2-T (1mer)	6.82±0.08	0.64±0.06	0.59±0.04	-1.76±0.04	2.34±0.06
DTT-T (1mer)	7.02±0.05	0.73±0.05	0.02±0.03	-2.43±0.03	2.45±0.05

Perhaps not surprisingly, the relative hydrophobicity of the overall compound mirrors the relative hydrophobicity of the constituent backbone comonomer. Meaning, the oligomer with a PDT backbone was the most hydrophobic (greatest log P and log D_{min}). The oligomer with the DTT backbone with the most hydrophilic (lowest log P and log D_{min}). And the oligomer with the amphipathic PEG2 backbone was in between. The pK_a's are close enough that all 3 should be protonated with similar kinetic profiles in an endosome. It's as if one could just translate one profile vertically to get one of the other two profiles. At least for the 1-mers, it suggests the possibility to decouple pK_a and hydrophobicity with regards to backbone choice because the backbone has a significant effect on global hydrophobicity but a minimal one on pK_a. In addition, increasing the hydrophobicity of the backbone caused a decrease in the Hill slope.



Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
PDT-T (2mer)	6.80±0.06	0.64±0.05	1.88±0.08	-2.72±0.07	4.60±0.11
PEG2-T (2mer)	7.61±0.04	1.33±0.13	0.61±0.04	-1.92±0.03	2.53±0.05
DTT-T (2mer)	8.08±0.04	1.22±0.09	-0.20±0.03	-2.05±0.01	1.85±0.03

As similarly seen in the 1-mers, increasing backbone hydrophobicity again increased the log P. However, with the addition of a 2nd mer, some new correlations arise. Increasing the hydrophobicity of the backbone resulted in decreasing the pK_a and increasing the span of log D.



Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
PDT-T (3mer)	7.18±0.01	1.55±0.12	2.06±0.02	-2.03±0.01	4.08±0.02
PEG2-T (3mer)	7.85±0.13	0.72±0.13	1.20±0.19	-3.13±0.12	4.33±0.24
DTT-T (3mer)	7.69±0.23	0.76±0.25	0.46±0.19	-2.09±0.13	2.54±0.24

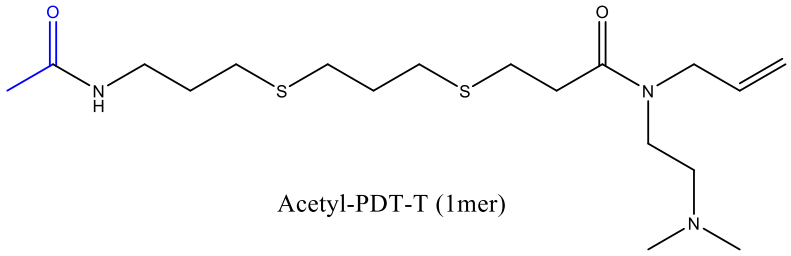
The family of 3-mers all show $pK_{a,\,average}$ that fall near or above physiological pH (i.e. 7.4), meaning all three will have some proportion of compound that is already ionized at physiological pH. The 3-mers also seem to have a larger span of log D than their 1- and 2- mer counterparts. For the 3-mers, increasing backbone hydrophobicity results in increasing log P, which is the same trend as the 1- and 2- mers.

Characterization of additional modifications on the library

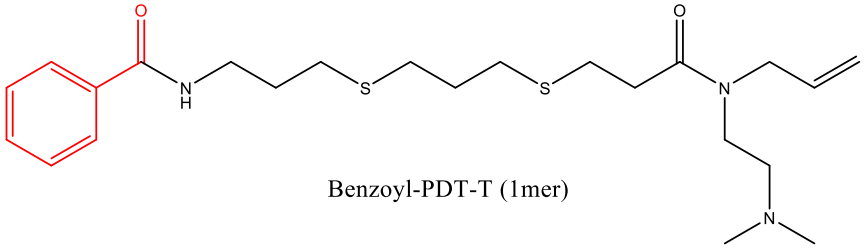
In addition to the original library, there were additional experiments done that looked at library members with additional modifications.

End Capping

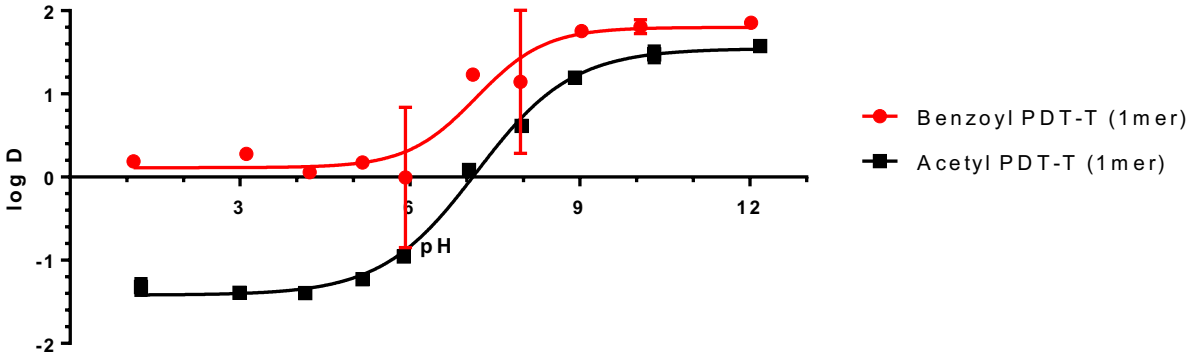
All members of the original library were capped with an acetyl group on the terminal primary amine (see left terminal of the first chemical structure shown below). A variant of the PDT-T (1mer) was made where the primary amine was instead capped with a benzoyl group.



Acetyl-PDT-T (1mer)



Benzoyl-PDT-T (1mer)



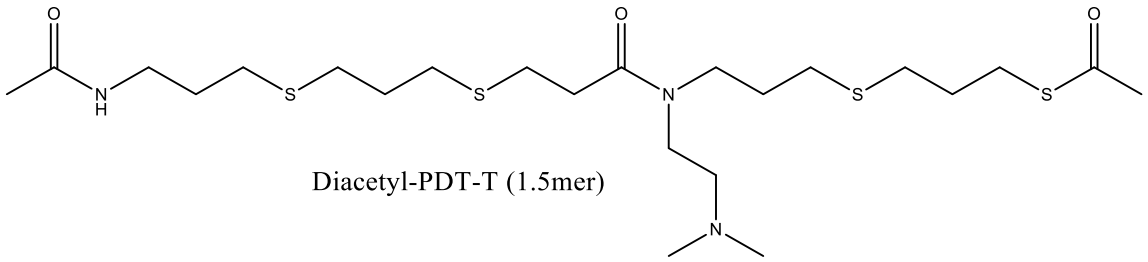
Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
Benzoyl-PDT-T (1mer)	7.17±0.32	0.71±0.34	1.80±0.15	0.11±0.13	1.69±0.21
Acetyl-PDT-T (1mer)	7.18±0.07	0.52±0.04	1.55±0.05	-1.42±0.04	2.97±0.07

It is interesting to note that they have nearly identical pK_a's but very different hydrophobicity profiles. This suggests two things. One, the local environment of the tertiary amine is far enough from the cap group topologically to be insensitive to the identity of the cap, which would explain why the pK_a doesn't change. Two, from a design perspective this comparison suggests it could be possible to modify hydrophobicity independently of the pK_a parameter based on the terminal conjugation.

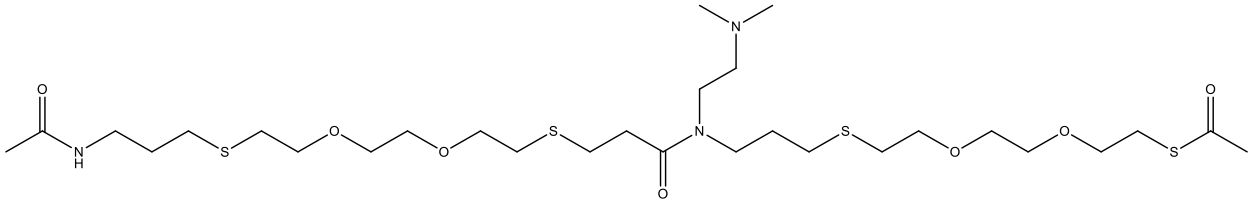
Half mers

In the original library, the addition of another dithiol comonomer was coupled to adding an additional possible charge to the oligomer (by only increasing the oligomer by integer mer units). Since one aspect of the study was to study how the backbone affects the local environment of the side chain, I made variants of some 1 mers. I added an additional 0.5 mer to see what the effect of an additional backbone unit would do.

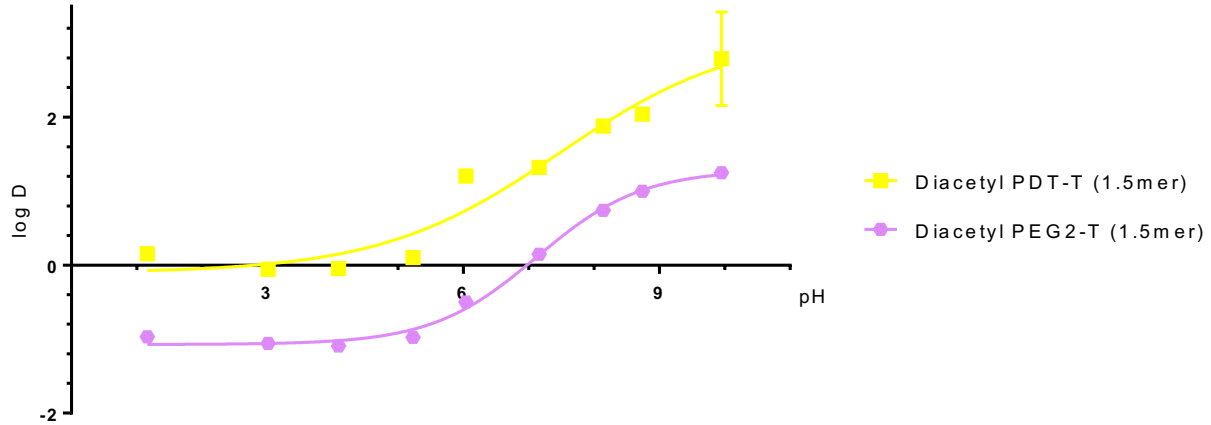
Due to the reactive nature of a free thiol, the terminal thiol of the additional dithiol comonomer is capped with an acetyl group. These two variants, diacetyl-PDT-T (1.5mer) and diacetyl-PEG2-T (1.5mer) are shown below.



Diacetyl-PDT-T (1.5mer)

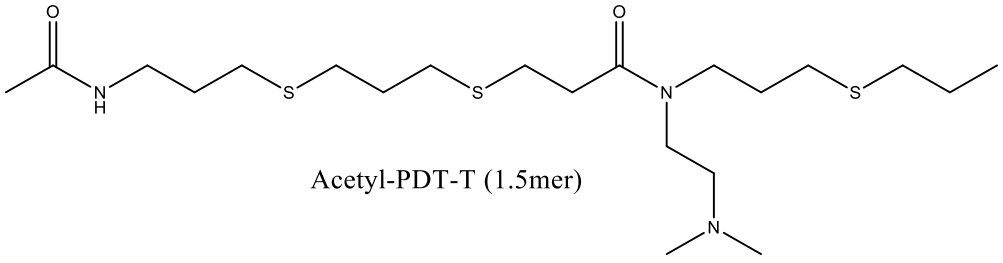


Diacetyl-PEG2-T (1.5mer)

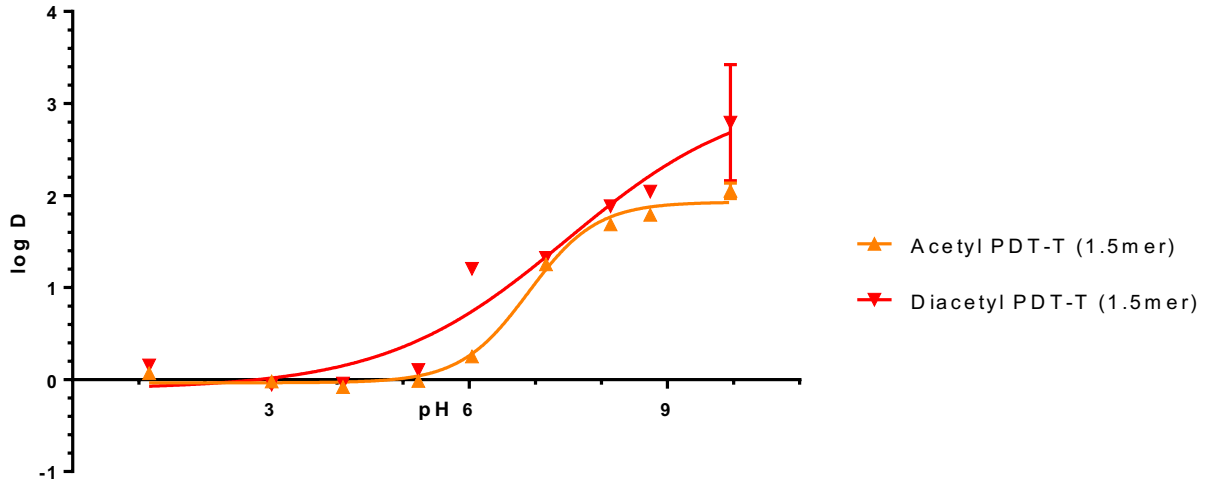


Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
Diacetyl-PDT-T (1.5mer)	7.48±0.67	0.32±0.12	3.15±0.68	-0.10±0.21	3.25±0.81
Diacetyl-PEG2-T (1.5mer)	7.12±0.08	0.54±0.05	1.29±0.07	-1.07±0.03	2.36±0.08

An additional variant of the PDT is made where a monothiol was added instead of the dithiol (thus a second cap is not necessary). This variant, acetyl-PDT-T (1.5mer), is shown below.



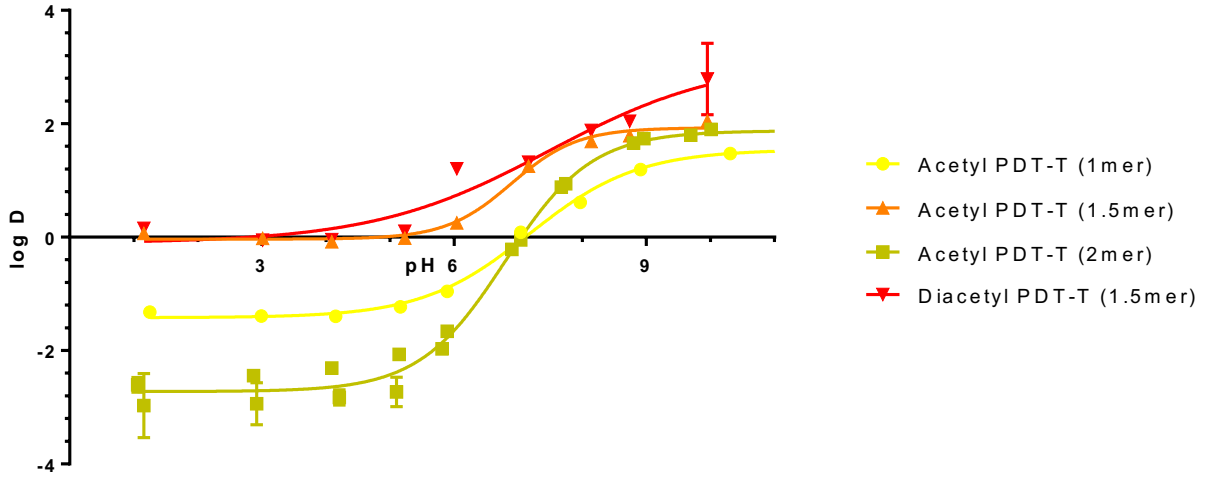
Acetyl-PDT-T (1.5mer)



Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
Acetyl-PDT-T (1.5mer)	6.90±0.08	0.84±0.10	1.93±0.05	-0.04±0.04	1.96±0.07
Diacetyl-PDT-T (1.5mer)	7.48±0.67	0.32±0.12	3.15±0.68	-0.10±0.21	3.25±0.81

The two variants of the PDT-T 1.5mer have fairly similar pK_a, but the diacetyl variant has a broader log D profile, as given by the span of log D.

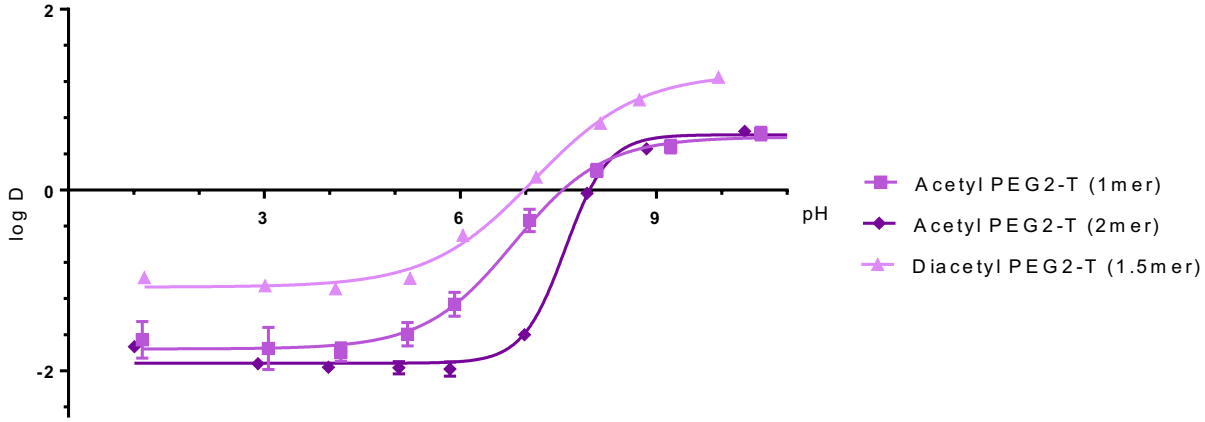
These variants can be plotted in the context of the adjacent integer mers in the synthesis to see what the addition of an additional half mer does.



Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
Acetyl-PDT-T (1mer)	7.18±0.07	0.52±0.04	1.55±0.05	-1.42±0.04	2.97±0.07

Acetyl-PDT-T (1.5mer)	6.90±0.08	0.84±0.10	1.93±0.05	-0.04±0.04	1.96±0.07
Acetyl-PDT-T (2mer)	6.80±0.06	0.64±0.05	1.88±0.08	-2.72±0.07	4.60±0.11
Diacetyl-PDT-T (1.5mer)	7.48±0.67	0.32±0.12	3.15±0.68	-0.10±0.21	3.25±0.81

For the single acetyl oligomers, increasing mer length caused a slight reduction in pK_a, but generally the pK_a for the 1.5 mers still fell within a similar range to the integer mers. The addition of another backbone group significantly increased the log D_{min} in both the 1.5 mers. And log P is significantly increased for the diacetyl case. With the addition of the second tertiary amine monomer, the hydrophobicity is significantly reduced at low pH.



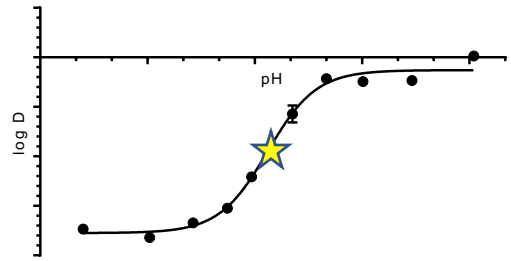
Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	Span of log D
Acetyl-PEG2-T (1mer)	6.82±0.08	0.64±0.06	0.59±0.04	-1.76±0.04	2.34±0.06
Diacetyl-PEG2-T (1.5mer)	7.12±0.08	0.54±0.05	1.29±0.07	-1.07±0.03	2.36±0.08
Acetyl-PEG2-T (2mer)	7.61±0.04	1.33±0.13	0.61±0.04	-1.92±0.03	2.53±0.05

The addition of diacetyl PEG2 backbone group leaves a curve with roughly the same span, except shifted in the hydrophobic direction. Both the log P and log D_{min} parameters were increased for the 1.5 mer. So in both the PDT and PEG2 family, the diacetyl addition results in a more hydrophobic log D_{min}. With both the PEG2 and PDT families, at low values of pH the hydrophobicity of the molecules decreased going from the 1-mer to the 1.5-mer, but then the hydrophobicity was rescued going from the 1.5-mer to the 2-mer. This perhaps hints that the terminal component plays a key role in solubility when the molecule is charged: namely that ending in a dithiol comonomer results in a more hydrophobic compound than molecules terminating with a tertiary amine monomer.

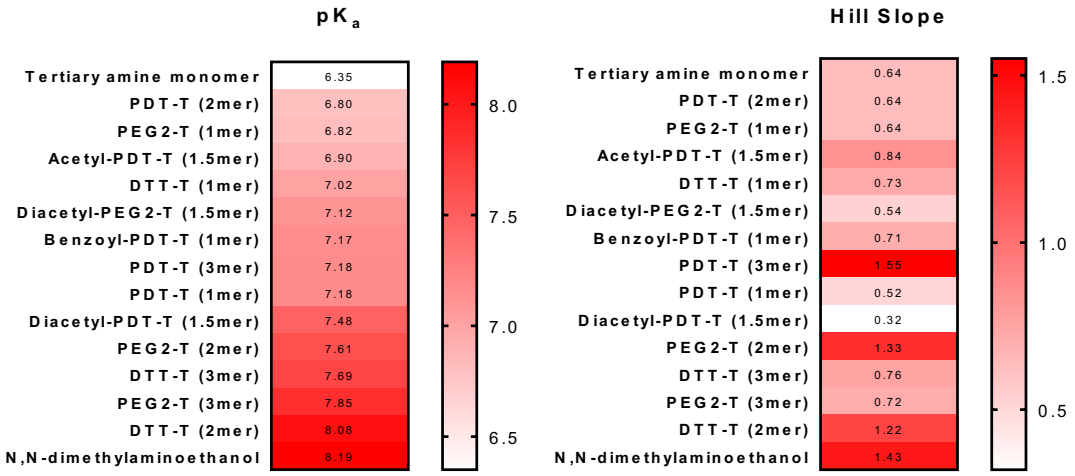
Comparison of parameters across all compounds

For each parameter, all compounds that have been tested were sorted in ascending order by parameter to see if there are any discernible trends. The results were ranked in a table so that the other parameters could be compared as well. A heatmap for each parameter is given so that potential clustering and outliers can be qualitatively visualized.

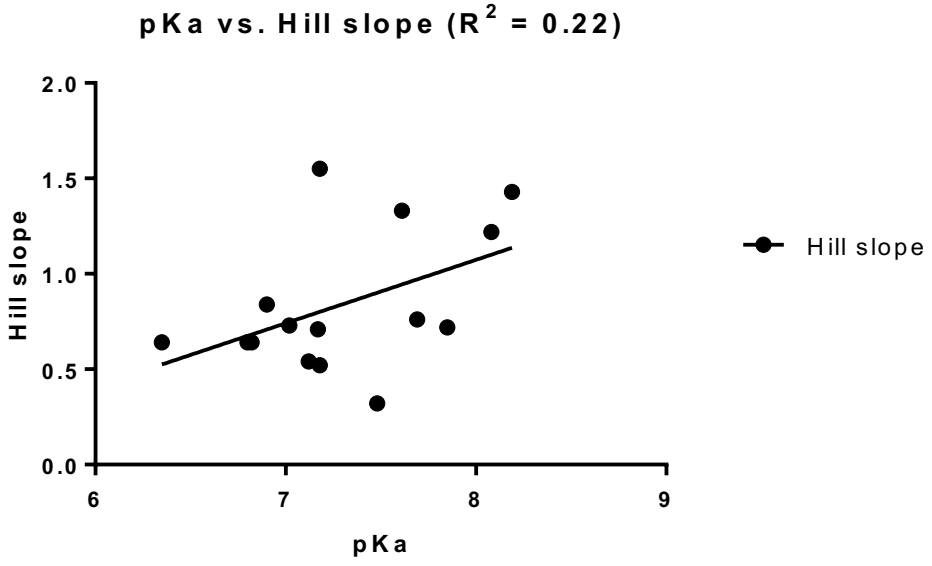
pK$_a$



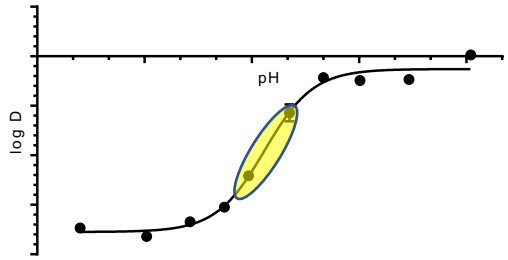
Compound	**pK$_a$**	Hill slope	log D$_{max}$ (log P)	log D$_{min}$	Span of log D
Tertiary amine monomer	**6.35±0.09**	0.64±0.07	-0.13±0.03	-1.77±0.04	1.65±0.05
PDT-T (2mer)	**6.80±0.06**	0.64±0.05	1.88±0.08	-2.72±0.07	4.60±0.11
PEG2-T (1mer)	**6.82±0.08**	0.64±0.06	0.59±0.04	-1.76±0.04	2.34±0.06
Acetyl-PDT-T (1.5mer)	**6.90±0.08**	0.84±0.10	1.93±0.05	-0.04±0.04	1.96±0.07
DTT-T (1mer)	**7.02±0.05**	0.73±0.05	0.02±0.03	-2.43±0.03	2.45±0.05
Diacetyl-PEG2-T (1.5mer)	**7.12±0.08**	0.54±0.05	1.29±0.07	-1.07±0.03	2.36±0.08
Benzoyl-PDT-T (1mer)	**7.17±0.32**	0.71±0.34	1.80±0.15	0.11±0.13	1.69±0.21
PDT-T (3mer)	**7.18±0.01**	1.55±0.12	2.06±0.02	-2.03±0.01	4.08±0.02
PDT-T (1mer)	**7.18±0.07**	0.52±0.04	1.55±0.05	-1.42±0.04	2.97±0.07
Diacetyl-PDT-T (1.5mer)	**7.48±0.67**	0.32±0.12	3.15±0.68	-0.10±0.21	3.25±0.81
PEG2-T (2mer)	**7.61±0.04**	1.33±0.13	0.61±0.04	-1.92±0.03	2.53±0.05
DTT-T (3mer)	**7.69±0.23**	0.76±0.25	0.46±0.19	-2.09±0.13	2.54±0.24
PEG2-T (3mer)	**7.85±0.13**	0.72±0.13	1.20±0.19	-3.13±0.12	4.33±0.24
DTT-T (2mer)	**8.08±0.04**	1.22±0.09	-0.20±0.03	-2.05±0.01	1.85±0.03
N,N-dimethylaminoethanol	**8.19±0.09**	1.43±0.47	-0.11±0.08	-1.48±0.04	1.37±0.09



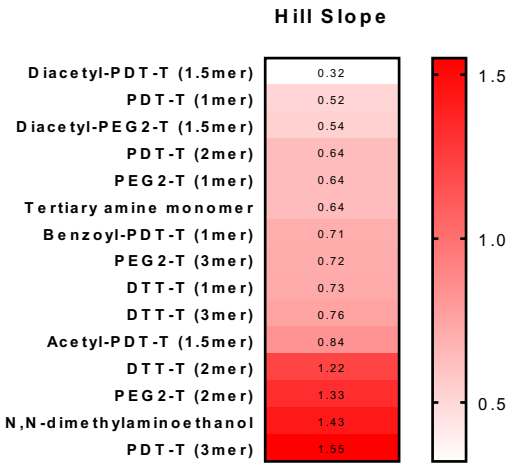
It is interesting to see from the heatmap that the tertiary amine monomer stands out as having a significantly lower pK_a than the rest of the library. It is also interesting that every oligomer has a pK_a in between the monomer and its subunit proxy (N,N-dimethylaminoethanol). Nearly all the compounds containing PDT are clustered fairly close (from 6.8 to 7.18). As seen in the regression plot below, there is a weak positive qualitative correlation of the pK_a parameter with the Hill slope parameter (R^2=0.22). Regressions for all pairs of parameters (plus an additional parameter, the molecular mass) are found in the 2-D Regression Plots between Parameters section (Supplementary Information).

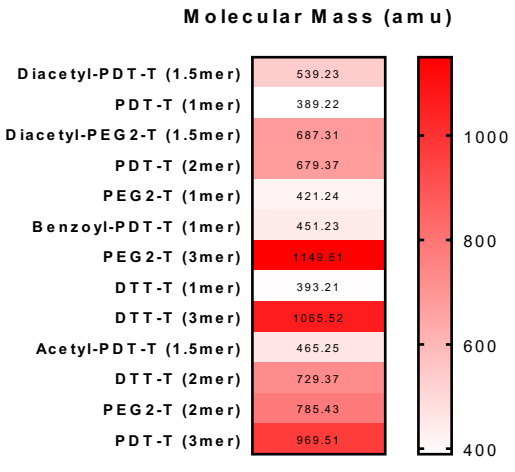


Hill slope

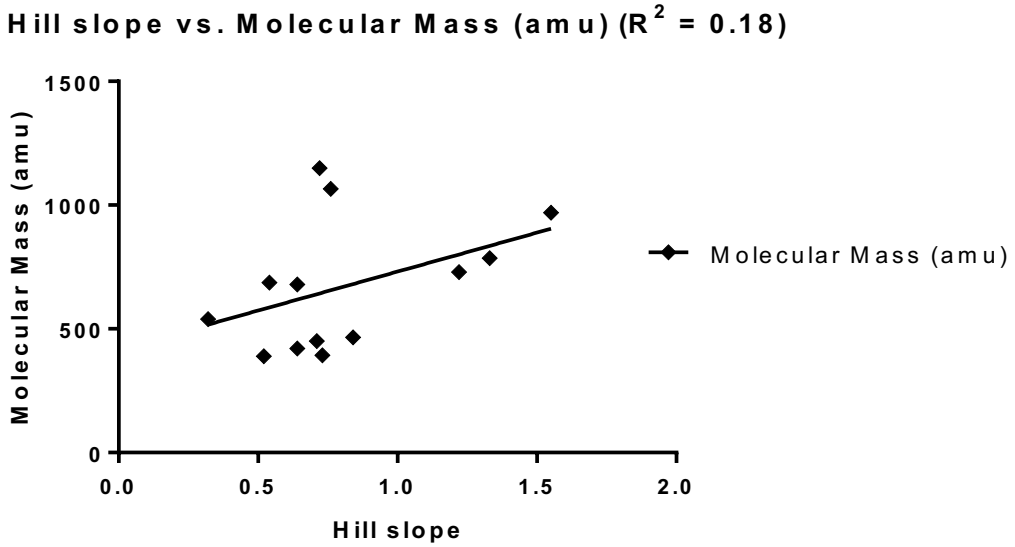


Compound	pK_a	**Hill slope**	log D_{max} (log P)	log D_{min}	Span of log D
Diacetyl-PDT-T (1.5mer)	7.48±0.67	**0.32±0.12**	3.15±0.68	-0.10±0.21	3.25±0.81
PDT-T (1mer)	7.18±0.07	**0.52±0.04**	1.55±0.05	-1.42±0.04	2.97±0.07
Diacetyl-PEG2-T (1.5mer)	7.12±0.08	**0.54±0.05**	1.29±0.07	-1.07±0.03	2.36±0.08
PDT-T (2mer)	6.80±0.06	**0.64±0.05**	1.88±0.08	-2.72±0.07	4.60±0.11
PEG2-T (1mer)	6.82±0.08	**0.64±0.06**	0.59±0.04	-1.76±0.04	2.34±0.06
Tertiary amine monomer	6.35±0.09	**0.64±0.07**	-0.13±0.03	-1.77±0.04	1.65±0.05
Benzoyl-PDT-T (1mer)	7.17±0.32	**0.71±0.34**	1.80±0.15	0.11±0.13	1.69±0.21
PEG2-T (3mer)	7.85±0.13	**0.72±0.13**	1.20±0.19	-3.13±0.12	4.33±0.24
DTT-T (1mer)	7.02±0.05	**0.73±0.05**	0.02±0.03	-2.43±0.03	2.45±0.05
DTT-T (3mer)	7.69±0.23	**0.76±0.25**	0.46±0.19	-2.09±0.13	2.54±0.24
Acetyl-PDT-T (1.5mer)	6.90±0.08	**0.84±0.10**	1.93±0.05	-0.04±0.04	1.96±0.07
DTT-T (2mer)	8.08±0.04	**1.22±0.09**	-0.20±0.03	-2.05±0.01	1.85±0.03
PEG2-T (2mer)	7.61±0.04	**1.33±0.13**	0.61±0.04	-1.92±0.03	2.53±0.05
N,N-dimethylaminoethanol	8.19±0.09	**1.43±0.47**	-0.11±0.08	-1.48±0.04	1.37±0.09
PDT-T (3mer)	7.18±0.01	**1.55±0.12**	2.06±0.02	-2.03±0.01	4.08±0.02

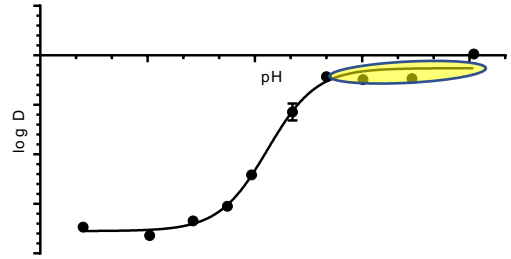




The heatmap indicates there are two primary populations for the Hill slope: one that contains slopes from 0.52 to 0.84 and another that contains slopes from 1.22 to 1.55. Within these two populations though, it is not quite clear if there are any trends in terms of sequence. There is a weak correlation between Hill slope and molecular mass (R^2 = 0.18). The molecular mass heatmap is included for comparison (mass of monomer and its subunit excluded for more resolution on scale).

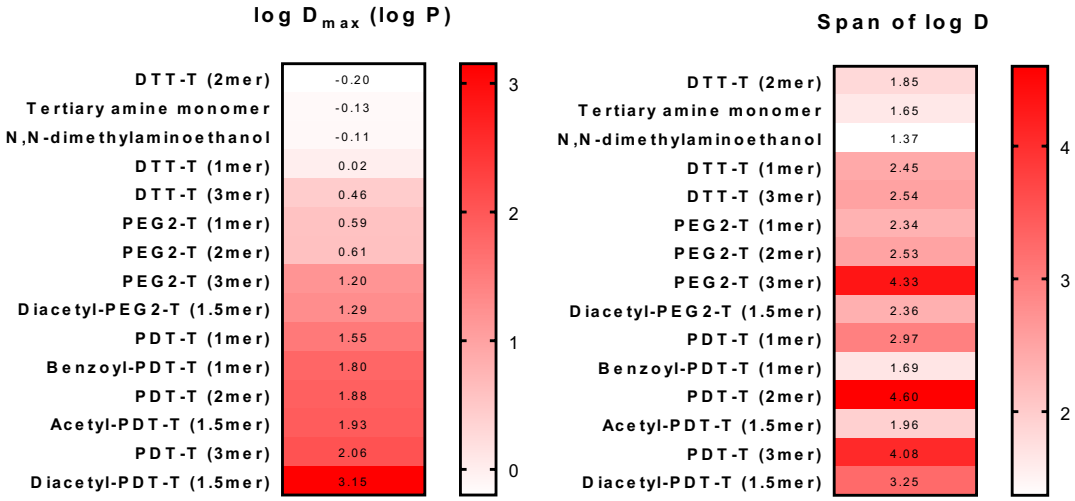


log D_{max} (log P)



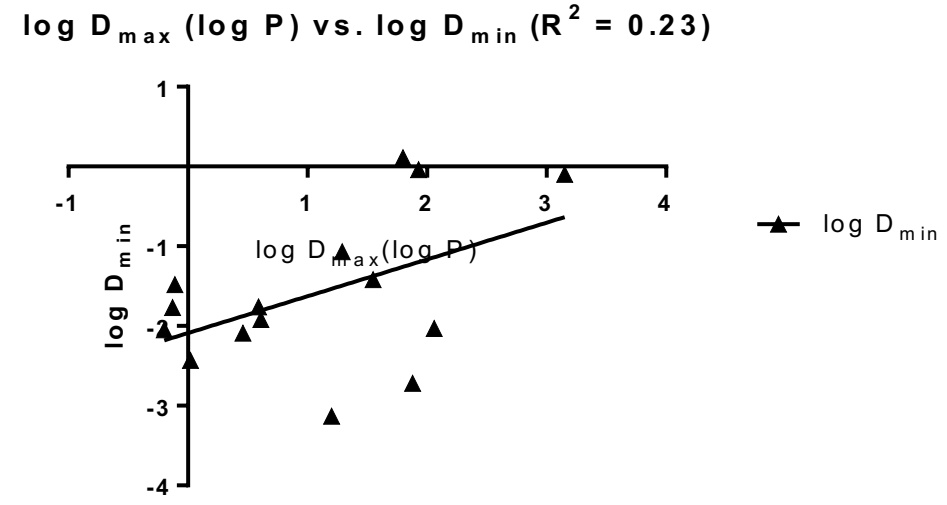
Compound	pK_a	Hill slope	**log D_{max} (log P)**	log D_{min}	Span of log D
DTT-T (2mer)	8.08±0.04	1.22±0.09	**-0.20±0.03**	-2.05±0.01	1.85±0.03
Tertiary amine monomer	6.35±0.09	0.64±0.07	**-0.13±0.03**	-1.77±0.04	1.65±0.05
N,N-dimethylaminoethanol	8.19±0.09	1.43±0.47	**-0.11±0.08**	-1.48±0.04	1.37±0.09
DTT-T (1mer)	7.02±0.05	0.73±0.05	**0.02±0.03**	-2.43±0.03	2.45±0.05
DTT-T (3mer)	7.69±0.23	0.76±0.25	**0.46±0.19**	-2.09±0.13	2.54±0.24
PEG2-T (1mer)	6.82±0.08	0.64±0.06	**0.59±0.04**	-1.76±0.04	2.34±0.06
PEG2-T (2mer)	7.61±0.04	1.33±0.13	**0.61±0.04**	-1.92±0.03	2.53±0.05

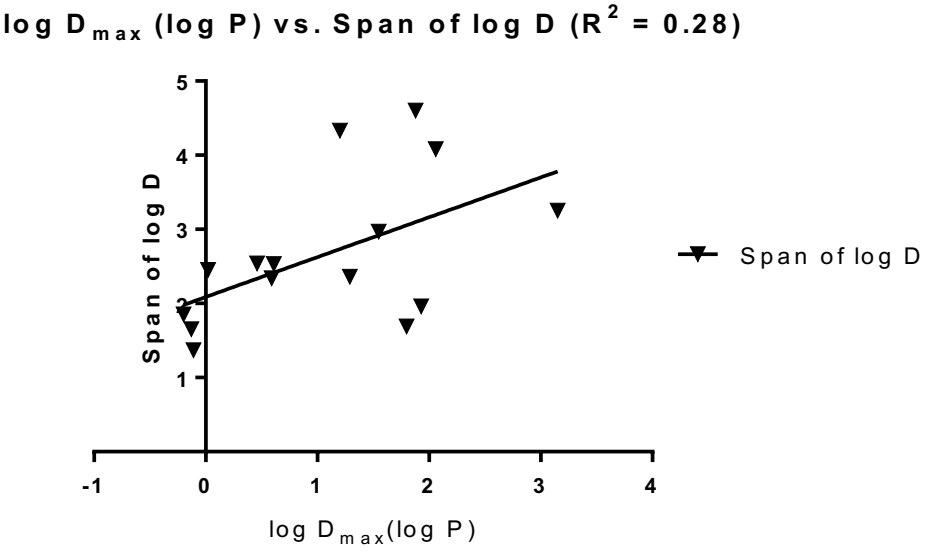
PEG2-T (3mer)	7.85±0.13	0.72±0.13	**1.20±0.19**	-3.13±0.12	4.33±0.24
Diacetyl-PEG2-T (1.5mer)	7.12±0.08	0.54±0.05	**1.29±0.07**	-1.07±0.03	2.36±0.08
PDT-T (1mer)	7.18±0.07	0.52±0.04	**1.55±0.05**	-1.42±0.04	2.97±0.07
Benzoyl-PDT-T (1mer)	7.17±0.32	0.71±0.34	**1.80±0.15**	0.11±0.13	1.69±0.21
PDT-T (2mer)	6.80±0.06	0.64±0.05	**1.88±0.08**	-2.72±0.07	4.60±0.11
Acetyl-PDT-T (1.5mer)	6.90±0.08	0.84±0.10	**1.93±0.05**	-0.04±0.04	1.96±0.07
PDT-T (3mer)	7.18±0.01	1.55±0.12	**2.06±0.02**	-2.03±0.01	4.08±0.02
Diacetyl-PDT-T (1.5mer)	7.48±0.67	0.32±0.12	**3.15±0.68**	-0.10±0.21	3.25±0.81



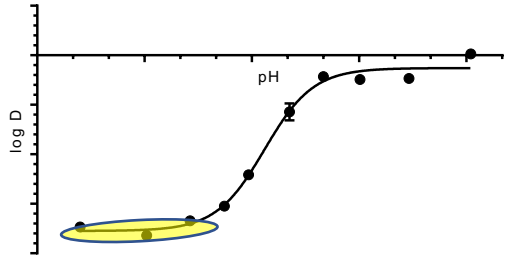
The log P values are nicely segregated based on backbone choice. DTT-containing compounds, the monomer, and the monomer subunit all have the lowest log P values (all hovering around 0 with the except of the DTT-T (3mer)). PEG2-containing compounds occupy a space slightly higher in log P values. And the PDT-containing compounds have the highest log P values (clustering around 1.5-2). Based on the relative hydrophobicities of these dithiols, the trends are expected. For the case of PEG2 and PDT compounds, increasing mer length also resulted in increased log P. The Diacetyl-PDT-T (1.5mer) also seems like an outlier based on the heatmap (with a log P greater than 1 unit over the next highest compound). There is also a small positive correlation of this parameter with

the two other hydrophobicity parameters: log D_{min} (R^2 = 0.23) and Span of log D (R^2 = 0.28).

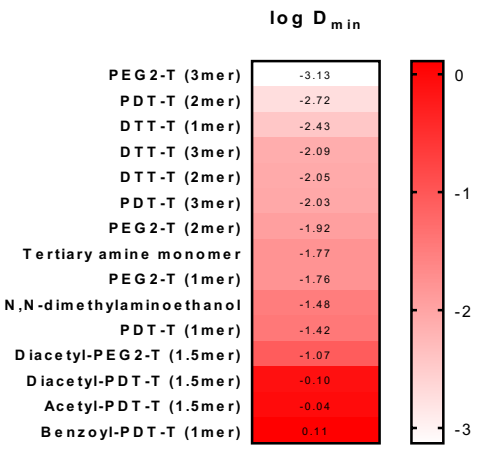


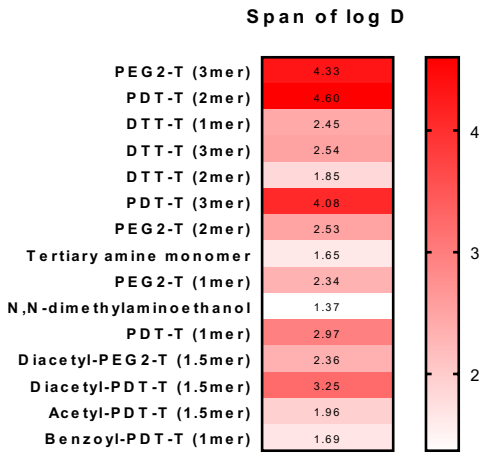


log D_{min}

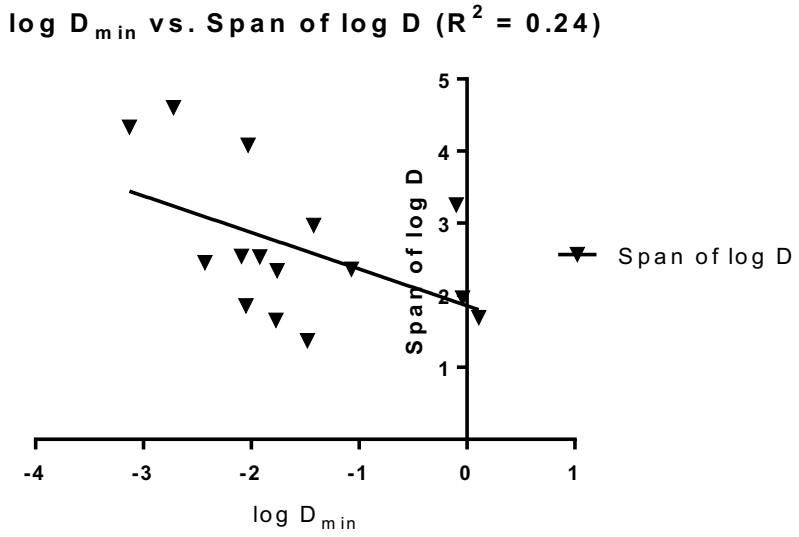


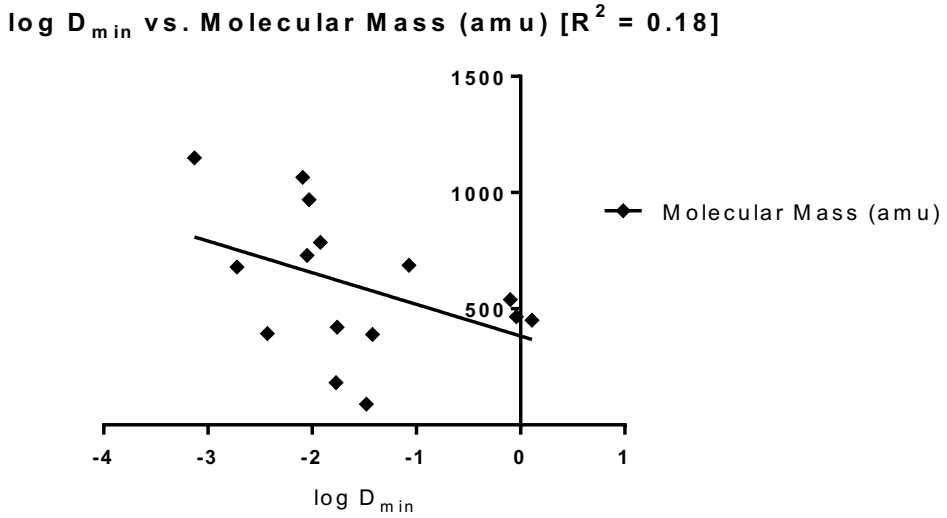
Compound	pK_a	Hill slope	log D_{max} (log P)	**log D_{min}**	Span of log D
PEG2-T (3mer)	7.85±0.13	0.72±0.13	1.20±0.19	**-3.13±0.12**	4.33±0.24
PDT-T (2mer)	6.80±0.06	0.64±0.05	1.88±0.08	**-2.72±0.07**	4.60±0.11
DTT-T (1mer)	7.02±0.05	0.73±0.05	0.02±0.03	**-2.43±0.03**	2.45±0.05
DTT-T (3mer)	7.69±0.23	0.76±0.25	0.46±0.19	**-2.09±0.13**	2.54±0.24
DTT-T (2mer)	8.08±0.04	1.22±0.09	-0.20±0.03	**-2.05±0.01**	1.85±0.03
PDT-T (3mer)	7.18±0.01	1.55±0.12	2.06±0.02	**-2.03±0.01**	4.08±0.02
PEG2-T (2mer)	7.61±0.04	1.33±0.13	0.61±0.04	**-1.92±0.03**	2.53±0.05
Tertiary amine monomer	6.35±0.09	0.64±0.07	-0.13±0.03	**-1.77±0.04**	1.65±0.05
PEG2-T (1mer)	6.82±0.08	0.64±0.06	0.59±0.04	**-1.76±0.04**	2.34±0.06
N,N-dimethylaminoethanol	8.19±0.09	1.43±0.47	-0.11±0.08	**-1.48±0.04**	1.37±0.09
PDT-T (1mer)	7.18±0.07	0.52±0.04	1.55±0.05	**-1.42±0.04**	2.97±0.07
Diacetyl-PEG2-T (1.5mer)	7.12±0.08	0.54±0.05	1.29±0.07	**-1.07±0.03**	2.36±0.08
Diacetyl-PDT-T (1.5mer)	7.48±0.67	0.32±0.12	3.15±0.68	**-0.10±0.21**	3.25±0.81
Acetyl-PDT-T (1.5mer)	6.90±0.08	0.84±0.10	1.93±0.05	**-0.04±0.04**	1.96±0.07
Benzoyl-PDT-T (1mer)	7.17±0.32	0.71±0.34	1.80±0.15	**0.11±0.13**	1.69±0.21



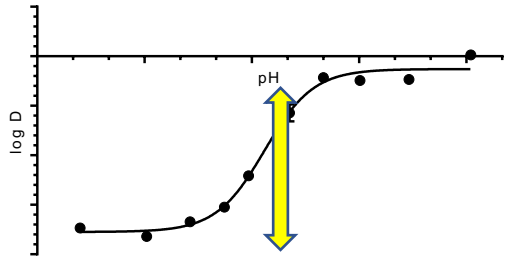


For most of the compounds, the log D_{min} parameter is segregated in a fashion similar to the log P values. The DTT-containing compounds are clustered from -2.43 to -2.05. The 1- and 2-mer of the PEG2 fall in the middle of the heat map. The majority of the compounds fall within the middle third of the heat map (-1 to -2). Surprisingly, the 3-mer of the PEG2 is an outlier from the other PEG2-containing compounds, with the smallest log D_{min} in the library. The variants of the PDT-T are outliers on the other end of the spectrum, clustering around 0. There is a weak inverse correlation of this parameter with both the Span of log D ($R^2 = 0.24$) and Molecular Mass ($R^2 = 0.18$).

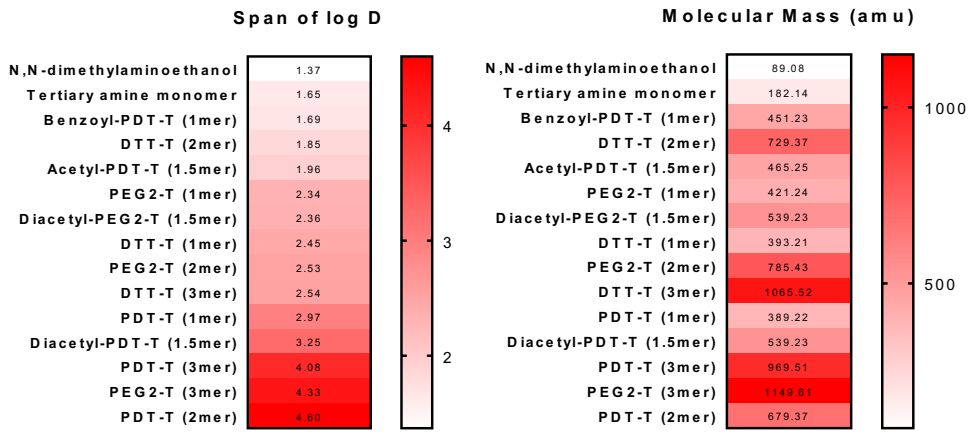




Span of log D

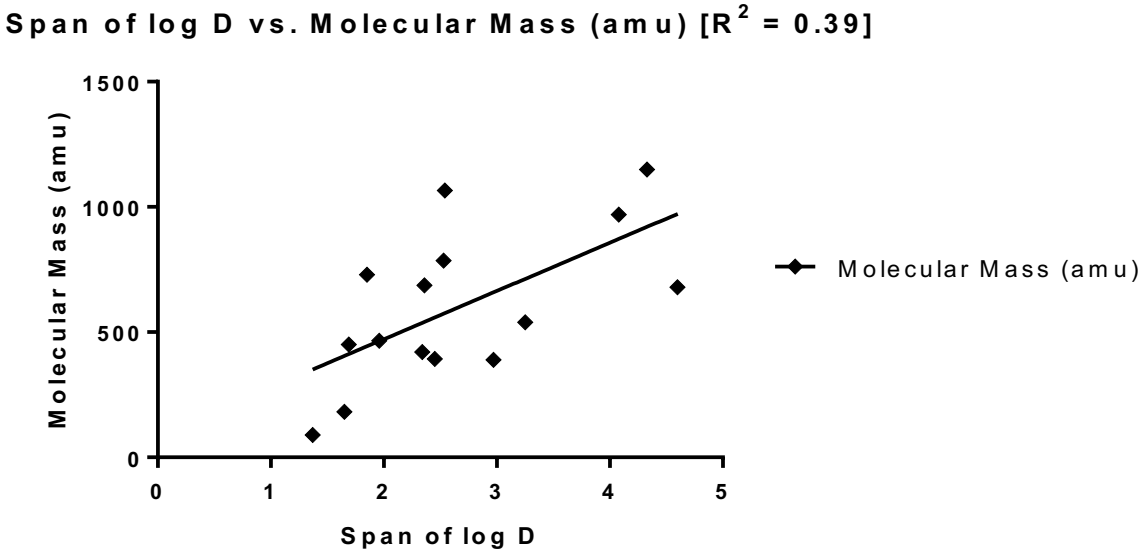


Compound	pK_a	Hill slope	log D_{max} (log P)	log D_{min}	**Span of log D**
N,N-dimethylaminoethanol	8.19±0.09	1.43±0.47	-0.11±0.08	-1.48±0.04	**1.37±0.09**
Tertiary amine monomer	6.35±0.09	0.64±0.07	-0.13±0.03	-1.77±0.04	**1.65±0.05**
Benzoyl-PDT-T (1mer)	7.17±0.32	0.71±0.34	1.80±0.15	0.11±0.13	**1.69±0.21**
DTT-T (2mer)	8.08±0.04	1.22±0.09	-0.20±0.03	-2.05±0.01	**1.85±0.03**
Acetyl-PDT-T (1.5mer)	6.90±0.08	0.84±0.10	1.93±0.05	-0.04±0.04	**1.96±0.07**
PEG2-T (1mer)	6.82±0.08	0.64±0.06	0.59±0.04	-1.76±0.04	**2.34±0.06**
Diacetyl-PEG2-T (1.5mer)	7.12±0.08	0.54±0.05	1.29±0.07	-1.07±0.03	**2.36±0.08**
DTT-T (1mer)	7.02±0.05	0.73±0.05	0.02±0.03	-2.43±0.03	**2.45±0.05**
PEG2-T (2mer)	7.61±0.04	1.33±0.13	0.61±0.04	-1.92±0.03	**2.53±0.05**
DTT-T (3mer)	7.69±0.23	0.76±0.25	0.46±0.19	-2.09±0.13	**2.54±0.24**
PDT-T (1mer)	7.18±0.07	0.52±0.04	1.55±0.05	-1.42±0.04	**2.97±0.07**
Diacetyl-PDT-T (1.5mer)	7.48±0.67	0.32±0.12	3.15±0.68	-0.10±0.21	**3.25±0.81**
PDT-T (3mer)	7.18±0.01	1.55±0.12	2.06±0.02	-2.03±0.01	**4.08±0.02**
PEG2-T (3mer)	7.85±0.13	0.72±0.13	1.20±0.19	-3.13±0.12	**4.33±0.24**
PDT-T (2mer)	6.80±0.06	0.64±0.05	1.88±0.08	-2.72±0.07	**4.60±0.11**



Of all the correlations between the possible combinations of parameters, the strongest one is between the Span of log D and Molecular Mass (R^2 = 0.39). Smaller compounds seem to dominate

the upper half of the list and larger compounds occupying the lower half. The library is segregated into 3 main populations. The first is from 1.65 to 1.96 with no clear relation between its members. The second population is from 2.34 to 2.54, containing PEG2 and DTT members only. In the 3rd group, many of the PDT family are clustered at the higher end of the list (specifically 2- and 3- mers as outliers at 4.6 and 4.08 respectively).



Part 3. Inference of nanoscale intramolecular interactions between side-chains

Thus far, the log D curves give insight into interaction between the side-chain and backbone. For the various families and groupings of oligomers, there are some qualitative trends that link the pK_a (which reflects protonation of the side chain) and hydrophobicity (which varies with choice of backbone group). We have not made any comment yet on the interaction between side-chains because the data from those curves are a bulk measurement. Again, this means that the data is blind to the distribution of ionization states in each solvent phase at each point on the log D curve. In this section, I discuss how equilibrium statistical mechanical models can be used to extract parameters that characterize those distributions. Then with knowledge of those parameters in hand, I then comment on interactions between side-chains that may be at play.

Comparison of Different Protonation Models

The overall objective is to model the equilibrium partitioning of a compound between an aqueous phase and an octanol phase. I start with an analysis of models for 3 possible protonation sites to illustrate the different models that might be used and what the appropriate choice is. Then I move to the case of 2- and 1- protonation site(s). I then test for fitting dependence of models based on the number of protonation sites. Finally, I use the fitted parameters from our models to interpret differences in the log D distributions for the compounds in our library.

Models for 3 Protonation Sites

The compound with 3 possible protonation sites corresponds to the 3 tertiary amines on the side-chains of an oligoTEA 3-mer. The reservoir of protons available is given by the pH of the aqueous phase. In the table below, I outline the differences between 3 simple models from which we might be able to calculate the distributions of ionization states.

Model Number	Interacting protonation sites?	Energetic difference in compound solvation energy between solvent phases?
Model 1	No	No
Model 2	No	Yes
Model 3	Yes	Yes

The models are different based on whether or not they make the assumption of (1) interacting protonation sites and (2) energetic difference in compound solvation energy between solvent phases. If the interacting protonation sites assumption is made, I chose the interaction to be modeled following a Pauling interaction model where all interactions are pairwise-symmetric. More elaborate models where higher-order interactions could be made, but I do not consider them here. If there is no energetic difference in compound solvation energy between the two solvent phases, termed symmetric solvent phases, then there is no energy difference associated with moving from the water to octanol phase. If there is a non-trivial energy difference, the compound has a different solvation energy in one phase compared to the other. Given the similarity of these models to Monod-Wyman-Changeux (MWC) models famous in biochemistry literature (cite Monod 1978), I also refer to this as a MWC-type difference. This energetic difference is analogous to the energetic difference between a tense and a relaxed state in an allosteric receptor. For those not so familiar with that literature, a simpler analogy can be considered as follows. Imagine your body being in two states: either sitting down on a chair or standing on your two feet. There is an energy associated with each state, where for instance, the standing state is the higher energy state because it takes more effort to be standing up than it does to remain seated. Now imagine I give you a cane. Having this cane alters the energy associated with both the sitting and standing states. It would take significantly less energy to be in the standing state if you have a cane, whereas the cane might only marginally decrease the energy associated with the sitting state. This change in the energy associated with both states is the MWC-type difference I refer to. The key power of the MWC model is that it provides a mechanism by which it is possible that changes to the system (analogous to the addition of the cane in our example here) can reverse the relative order of the energy of states (which could hypothetically make the sitting state higher in energy than the standing state)!

Using the grand canonical ensemble where the system is a 3-site compound and the reservoir is represented by the free protons dispersed in water and octanol, I use the Gibbs factor to assign weights to the possible states in each phase. Then, I use the equilibrium probabilities of each state to calculate the partitioning of the compound between the aqueous and octanol phases. Calculations are adapted from examples given by Phillips et. al[50] for hemoglobin modeling.

Model 1: Non-interacting protonation sites. No energetic difference in compound solvation energy between solvent phases (symmetric solvent phases)

Phase	Microstate	Protonation State	Weight
Octanol	_ _ _	0	1
Octanol	* _ _	1	$e^{-\beta(\epsilon_{oct}-\mu)}$
Octanol	_ * _	1	$e^{-\beta(\epsilon_{oct}-\mu)}$
Octanol	_ _ *	1	$e^{-\beta(\epsilon_{oct}-\mu)}$
Octanol	* * _	2	$e^{-\beta(2\epsilon_{oct}-2\mu)}$
Octanol	_ * *	2	$e^{-\beta(2\epsilon_{oct}-2\mu)}$
Octanol	* _ *	2	$e^{-\beta(2\epsilon_{oct}-2\mu)}$
Octanol	* * *	3	$e^{-\beta(3\epsilon_{oct}-3\mu)}$
Water	_ _ _	0	1
Water	* _ _	1	$e^{-\beta(\epsilon_{w}-\mu)}$
Water	_ * _	1	$e^{-\beta(\epsilon_{w}-\mu)}$
Water	_ _ *	1	$e^{-\beta(\epsilon_{w}-\mu)}$
Water	* * _	2	$e^{-\beta(2\epsilon_{w}-2\mu)}$
Water	_ * *	2	$e^{-\beta(2\epsilon_{w}-2\mu)}$
Water	* _ *	2	$e^{-\beta(2\epsilon_{w}-2\mu)}$
Water	* * *	3	$e^{-\beta(3\epsilon_{w}-3\mu)}$

Here, the state reflects which of the side-chains have been protonated, with the total number of side-chains protonated given by the protonation state. The symbols for the weights are defined as follows: $\beta \equiv \frac{1}{k_B T}$ is the inverse temperature, ϵ_{oct} is the energetic cost of adding 1 proton to the compound when the compound is in the octanol phase, ϵ_w is the energetic cost of adding 1 proton to the compound when the compound is in the aqueous phase, μ is the chemical potential associated with moving 1 proton from the reservoir into the system (i.e. protonating the macromolecule).

The weights are given in accordance with states defined in a grand canonical ensemble. A short explanation is given as follows. Consider a system and a surrounding reservoir that are isolated but allowed to exchange heat and particles with each other. A state then refers to an ensemble containing microstates that are indistinguishable from each other according to some property (ex. here that property is the protonation state). A fundamental postulate of statistical thermodynamics states that each microstate within a state has equal probability of being occupied. It follows then that

the probability of observing a system in a particular state *i*, $P_{system}(state_i)$, is proportional to the multiplicity of that state, $W_{system}(state_i)$:

$$P_{system}(state_i), \propto W_{system}(state_i)$$

The multiplicity is related to a quantity known as the entropy through Boltzmann's equation:

$$S = k \ln W$$

If the system has a multiplicity, W_{system}, and the reservoir, $W_{reservoir}$, then:

$$W_{total} = W_{system} \cdot W_{reservoir}$$

And by extension:

$$S_{total} = S_{system} + S_{reservoir}$$

Since we make the specification that the system is in a particular state *i*, then $W_{system} = 1$ and $S_{system} = 0$. It turns out then that the probability of state *i* is a function of $S_{reservoir}$.

$$P_{system}(state_i), \propto e^{\frac{S_{reservoir}(state_i)}{k}}$$

Using the fundamental thermodynamic relation, $dS = \left(\frac{1}{T}\right)(dE + PdV + -\mu dN)$,

$$P_{system}(state_i), \propto e^{\left(\frac{1}{kT_R}\right)(dE_R + PdV_R - \mu dN_R)}$$

In the grand canonical ensemble, the chemical potential (μ), the volume (V), and temperature (T) are constant, which reduces the expression to:

$$P_{system}(state_i), \propto e^{\left(\frac{1}{kT_R}\right)(dE_R - \mu dN_R)}$$

Since the total energy and particles between the system and reservoir are conserved and the temperature of the system and reservoir will be equal at equilibrium,

System Energy (E)+ Reservoir Energy (E_R) = Total Energy (E_T)

System Particle Number (N) + Reservoir Particle Number (N_R) = Total Particle Number (N_T)

System Temperature (T) = Reservoir Temperature (T_R)

one can make substitutions to put the expression in terms of system variables.

$$P_{system}(state_i), \propto e^{\left(\frac{-1}{kT}\right)(dE - \mu dN)}$$

The weight for a state is simply equal to the quantity on the right-hand side specific for that state.

$$Weight(state_i) = e^{-\beta(dE_i - \mu dN_i)}$$

With the weights for all the states defined, that directly allows one to write expressions for the probability of each state. Therefore, I can write the distribution coefficient in terms of the probabilities of all the protonation states of the compound in the octanol phases and the probabilities of all the protonation states of the compound in the aqueous phase. Here, I derive an expression of D that can be fit using the experimental data.

$$D \equiv \text{Distribution coefficient} \equiv \frac{\text{Equilibrium concentration of all states in the octanol phase}}{\text{Equilibrium concentration of all states in the water phase}}$$

$$D = \frac{\sum_i P_{octanol}, i}{\sum_j P_{water}, j}$$

$$= \frac{(1 + e^{-\beta(\epsilon_{oct}-\mu)} + e^{-\beta(\epsilon_{oct}-\mu)} + e^{-\beta(\epsilon_{oct}-\mu)} + e^{-\beta(2\epsilon_{oct}-2\mu)} + e^{-\beta(2\epsilon_{oct}-2\mu)} + e^{-\beta(2\epsilon_{oct}-2\mu)} + e^{-\beta(3\epsilon_{oct}-3\mu)})/Z}{(1 + e^{-\beta(\epsilon_w-\mu)} + e^{-\beta(\epsilon_w-\mu)} + e^{-\beta(\epsilon_w-\mu)} + e^{-\beta(2\epsilon_w-2\mu)} + e^{-\beta(2\epsilon_w-2\mu)} + e^{-\beta(2\epsilon_w-2\mu)} + e^{-\beta(3\epsilon_w-3\mu)})/Z}$$

$$= \frac{1 + 3e^{-\beta(\epsilon_{oct}-\mu)} + 3e^{-\beta(2\epsilon_{oct}-2\mu)} + e^{-\beta(3\epsilon_{oct}-3\mu)}}{1 + 3e^{-\beta(\epsilon_w-\mu)} + 3e^{-\beta(2\epsilon_w-2\mu)} + e^{-\beta(3\epsilon_w-3\mu)}}$$

For dilute solutions (these experiments are sub mM), I can use the substitution $\mu = \mu_0 + \frac{\ln\left(\frac{c}{c_0}\right)}{\beta}$:

$$D = \frac{1+3e^{-\beta\left(\epsilon_{oct}-(\mu_0+\frac{\ln\left(\frac{c}{c_0}\right)}{\beta})\right)}+3e^{-\beta\left(2\epsilon_{oct}-2(\mu_0+\frac{\ln\left(\frac{c}{c_0}\right)}{\beta})\right)}+e^{-\beta\left(3\epsilon_{oct}-3(\mu_0+\frac{\ln\left(\frac{c}{c_0}\right)}{\beta})\right)}}{1+3e^{-\beta\left(\epsilon_w-(\mu_0+\frac{\ln\left(\frac{c}{c_0}\right)}{\beta})\right)}+3e^{-\beta\left(2\epsilon_w-2(\mu_0+\frac{\ln\left(\frac{c}{c_0}\right)}{\beta})\right)}+e^{-\beta\left(3\epsilon_w-3(\mu_0+\frac{\ln\left(\frac{c}{c_0}\right)}{\beta})\right)}}$$

$$= \frac{1 + 3e^{-\beta\epsilon_{oct}+\beta\mu_0+\ln\left(\frac{c}{c_0}\right)} + 3e^{-2\beta\epsilon_{oct}+2\beta\mu_0+2\ln\left(\frac{c}{c_0}\right)} + e^{-3\beta\epsilon_{oct}+3\beta\mu_0+3\ln\left(\frac{c}{c_0}\right)}}{1 + 3e^{-\beta\epsilon_w+\beta\mu_0+\ln\left(\frac{c}{c_0}\right)} + 3e^{-2\beta\epsilon_w+2\beta\mu_0+2\ln\left(\frac{c}{c_0}\right)} + e^{-3\beta\epsilon_w+3\beta\mu_0+3\ln\left(\frac{c}{c_0}\right)}}$$

Using $c = [H^+]$ and defining ln with log:

$$= \frac{1 + 3e^{-\beta\epsilon_{oct}+\beta\mu_0+\frac{\log[H^+]}{\log(e)}-\ln(c_0)} + 3e^{-2\beta\epsilon_{oct}+2\beta\mu_0+2(\frac{\log[H^+]}{\log(e)}-\ln(c_0))} + e^{-3\beta\epsilon_{oct}+3\beta\mu_0+3(\frac{\log[H^+]}{\log(e)}-\ln(c_0))}}{1 + 3e^{-\beta\epsilon_w+\beta\mu_0+\frac{\log[H^+]}{\log(e)}-\ln(c_0)} + 3e^{-2\beta\epsilon_w+2\beta\mu_0+2(\frac{\log[H^+]}{\log(e)}-\ln(c_0))} + e^{-3\beta\epsilon_w+3\beta\mu_0+3(\frac{\log[H^+]}{\log(e)}-\ln(c_0))}}$$

Using $pH = -\log[H^+]$:

$$= \frac{1 + 3e^{-\beta\epsilon_{oct}+\beta\mu_0-\frac{pH}{\log(e)}-\ln(c_0)} + 3e^{-2\beta\epsilon_{oct}+2\beta\mu_0-2\frac{pH}{\log(e)}-2\ln(c_0)} + e^{-3\beta\epsilon_{oct}+3\beta\mu_0-3\frac{pH}{\log(e)}-3\ln(c_0)}}{1 + 3e^{-\beta\epsilon_w+\beta\mu_0-\frac{pH}{\log(e)}-\ln(c_0)} + 3e^{-2\beta\epsilon_w+2\beta\mu_0-2\frac{pH}{\log(e)}-2\ln(c_0)} + e^{-3\beta\epsilon_w+3\beta\mu_0-3\frac{pH}{\log(e)}-3\ln(c_0)}}$$

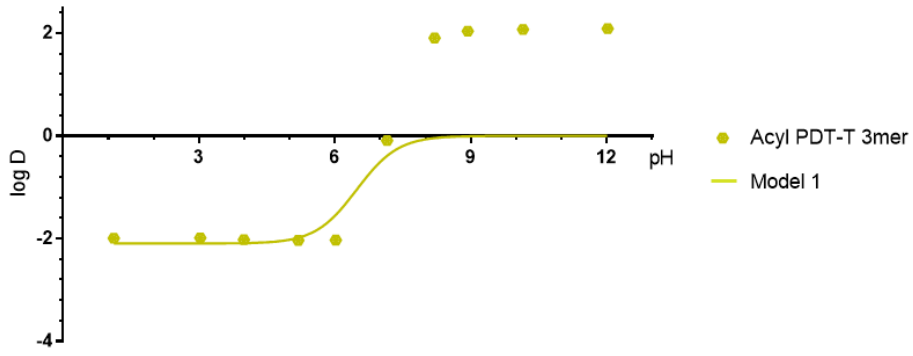
Collecting β terms and define: $oct \equiv \beta(\epsilon_{oct} - \mu_0)$ $water \equiv \beta(\epsilon_w - \mu_0)$

$$= \frac{1 + 3e^{-oct - \frac{\mathrm{pH}}{\log(e)} - \ln(c_0)} + 3e^{-2oct - 2\frac{\mathrm{pH}}{\log(e)} - 2\ln(c_0)} + e^{-3oct - 3\frac{\mathrm{pH}}{\log(e)} - 3\ln(c_0)}}{1 + 3e^{-water - \frac{\mathrm{pH}}{\log(e)} - \ln(c_0)} + 3e^{-2water - 2\frac{\mathrm{pH}}{\log(e)} - 2\ln(c_0)} + e^{-3water - 3\frac{\mathrm{pH}}{\log(e)} - 3\ln(c_0)}}$$

Taking log of both sides:

$$\log D = \log\left[\frac{1 + 3e^{-oct - \frac{pH}{log(e)} - ln(c_0)} + 3e^{-2oct - 2\frac{pH}{log(e)} - 2ln(c_0)} + e^{-3oct - 3\frac{pH}{log(e)} - 3ln(c_0)}}{1 + 3e^{-water - \frac{pH}{log(e)} - ln(c_0)} + 3e^{-2water - 2\frac{pH}{log(e)} - 2ln(c_0)} + e^{-3water - 3\frac{pH}{log(e)} - 3ln(c_0)}}\right]$$

Given experimental data of log D vs. pH, I fit for the parameters *oct* and *water*. The value c_0 was set to 10^{-7} because that is the concentration of protons in water at the neutral pH of 7, which is a natural reference state.



This above graph shows what happens if you try to fit this model with data from the PDT-T 3-mer. Clearly, this is not a good model. The model cannot produce log D values greater than 0. After inspecting Model 2, it will be apparent that the issue is due to assumption of symmetric solvent phases.

Model 2: Non-interacting protonation sites. Non-trivial energetic difference in compound solvation energy between solvent phases

Phase	Microstate	Protonation State	Weight
Octanol	$\underline{\quad}\ \underline{\quad}\ \underline{\quad}$	0	1
Octanol	$*\ \underline{\quad}\ \underline{\quad}\ \underline{\quad}$	1	$e^{-\beta(\epsilon_{oct} - \mu)}$
Octanol	$\underline{\quad}\ *\ \underline{\quad}\ \underline{\quad}$	1	$e^{-\beta(\epsilon_{oct} - \mu)}$
Octanol	$\underline{\quad}\ \underline{\quad}\ *\ \underline{\quad}$	1	$e^{-\beta(\epsilon_{oct} - \mu)}$
Octanol	$*\ *\ \underline{\quad}\ \underline{\quad}\ \underline{\quad}$	2	$e^{-\beta(2\epsilon_{oct} - 2\mu)}$

Octanol	$\_\ *\ *$	2	$e^{-\beta(2\epsilon_{oct}-2\mu)}$
Octanol	$*\ \_\ *$	2	$e^{-\beta(2\epsilon_{oct}-2\mu)}$
Octanol	$*\ *\ *$	3	$e^{-\beta(3\epsilon_{oct}-3\mu)}$
Water	$\_\ \_\ \_$	0	$e^{-\beta\epsilon}$
Water	$*\ \_\ \_$	1	$e^{-\beta(\epsilon_w-\mu+\epsilon)}$
Water	$\_\ *\ \_$	1	$e^{-\beta(\epsilon_w-\mu+\epsilon)}$
Water	$\_\ \_\ *$	1	$e^{-\beta(\epsilon_w-\mu+\epsilon)}$
Water	$*\ *\ \_$	2	$e^{-\beta(2\epsilon_w-2\mu+\epsilon)}$
Water	$\_\ *\ *$	2	$e^{-\beta(2\epsilon_w-2\mu+\epsilon)}$
Water	$*\ \_\ *$	2	$e^{-\beta(2\epsilon_w-2\mu+\epsilon)}$
Water	$*\ *\ *$	3	$e^{-\beta(3\epsilon_w-3\mu+\epsilon)}$

In comparison to Model 1, the key conceptual difference with Model 2 is the introduction of an energy cost (ϵ) of the compound being in the water phase (the energy cost of being in the octanol phase is defined for convenience as 0). This is more realistic than Model 1 because one would not expect the solvation energy of the compound in water to be the same as the solvation energy of the compound in octanol. This model is MWC-like because the protonation reaction has different energy costs depending on the solvent phase, which is analogous to the MWC model in enzymes where a ligand has a different binding affinity depending on the conformation of the enzyme. A positive ϵ indicates the solvation energy in the octanol phase is less than the solvation energy in the aqueous phase. Conversely, a negative ϵ indicates the solvation energy in the octanol phase is greater than the solvation energy in the aqueous phase. A calculation for D for Model 2 can be done following the same steps outlined above for Model 1.

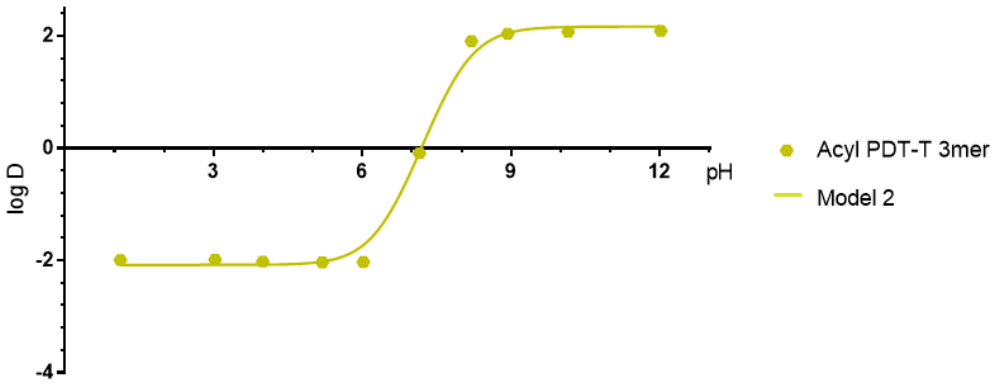
$$D = \frac{\sum_i P_{octanol}, i}{\sum_j P_{water}, j}$$

$$= \frac{(1 + e^{-\beta(\epsilon_{oct}-\mu)} + e^{-\beta(\epsilon_{oct}-\mu)} + e^{-\beta(\epsilon_{oct}-\mu)} + e^{-\beta(2\epsilon_{oct}-2\mu)} + e^{-\beta(2\epsilon_{oct}-2\mu)} + e^{-\beta(2\epsilon_{oct}-2\mu)} + e^{-\beta(3\epsilon_{oct}-3\mu)})/Z}{(e^{-\beta\epsilon} + e^{-\beta(\epsilon_w-\mu+\epsilon)} + e^{-\beta(\epsilon_w-\mu+\epsilon)} + e^{-\beta(\epsilon_w-\mu+\epsilon)} + e^{-\beta(2\epsilon_w-2\mu+\epsilon)} + e^{-\beta(2\epsilon_w-2\mu+\epsilon)} + e^{-\beta(2\epsilon_w-2\mu+\epsilon)} + e^{-\beta(3\epsilon_w-3\mu+\epsilon)})/Z}$$

$$= \frac{1 + 3e^{-\beta(\epsilon_{oct}-\mu)} + 3e^{-\beta(2\epsilon_{oct}-2\mu)} + e^{-\beta(3\epsilon_{oct}-3\mu)}}{e^{-\beta\epsilon} + 3e^{-\beta(\epsilon_w-\mu+\epsilon)} + 3e^{-\beta(2\epsilon_w-2\mu+\epsilon)} + e^{-\beta(3\epsilon_w-3\mu+\epsilon)}}$$

Following the same intermediate steps as Model 1, with this additional definition: $diff \equiv \beta\epsilon$

$$\log D = \log\left[\frac{1 + 3e^{-oct-\frac{pH}{log(e)}-ln(c_0)} + 3e^{-2oct-2\frac{pH}{log(e)}-2ln(c_0)} + e^{-3oct-3\frac{pH}{log(e)}-3\,ln(c_0)}}{e^{-diff} + 3e^{-water-diff-\frac{pH}{log(e)}-ln(c_0)} + 3e^{-2water-diff-2\frac{pH}{log(e)}-2ln(c_0)} + e^{-3water-diff-3\frac{pH}{log(e)}-3\,ln(c_0)}}\right]$$



The graph above is the same data set for the PDT-T 3-mer, now fitted with Model 2 for parameters *oct, water,* and *diff*. Thus, it is apparent that one requires an energetic difference between the two solvent phases to qualitatively recapitulate the experimental data. For the most part, this model seems to fit quite well. Flanking the inflection point, the model is too shallow and does not quite match the experimental data points corresponding to pH of 6 and 8.

Model 3: Interacting (Pauling model) protonation sites. Non-trivial energetic difference in compound solvation energy between solvent phases

Phase	Microstate	Protonation State	Number of Protonated Pairs	Weight
Octanol	_ _ _	0	0	1
Octanol	* _ _	1	0	$e^{-\beta(\epsilon_{oct}-\mu)}$
Octanol	_ * _	1	0	$e^{-\beta(\epsilon_{oct}-\mu)}$
Octanol	_ _ *	1	0	$e^{-\beta(\epsilon_{oct}-\mu)}$
Octanol	* * _	2	1	$e^{-\beta(2\epsilon_{oct}-2\mu+J)}$
Octanol	_ * *	2	1	$e^{-\beta(2\epsilon_{oct}-2\mu+J)}$
Octanol	* _ *	2	1	$e^{-\beta(2\epsilon_{oct}-2\mu+J)}$
Octanol	* * *	3	3	$e^{-\beta(3\epsilon_{oct}-3\mu+3J)}$
Water	_ _ _	0	0	$e^{-\beta\epsilon}$
Water	* _ _	1	0	$e^{-\beta(\epsilon_w-\mu+\epsilon)}$
Water	_ * _	1	0	$e^{-\beta(\epsilon_w-\mu+\epsilon)}$
Water	_ _ *	1	0	$e^{-\beta(\epsilon_w-\mu+\epsilon)}$
Water	* * _	2	1	$e^{-\beta(2\epsilon_w-2\mu+\epsilon+J)}$
Water	_ * *	2	1	$e^{-\beta(2\epsilon_w-2\mu+\epsilon+J)}$
Water	* _ *	2	1	$e^{-\beta(2\epsilon_w-2\mu+\epsilon+J)}$
Water	* * *	3	3	$e^{-\beta(3\epsilon_w-3\mu+\epsilon+3J)}$

In comparison with Model 2, the key conceptual difference with Model 3 is the introduction of interacting sites. Borrowing terminology again from biochemistry, the presence of interaction is given the name cooperativity. There is an energy difference (J) associated with protonating the compound if it has already been protonated. In this model, I only consider pairwise cooperativity (a Pauling interaction model), which means there is an additional amount of energy J for each pair of protonated sites in the state. This assumes all pairwise interactions among sites are symmetric in energy. Here a positive J indicates negative cooperativity, meaning existing protonated sites make it more difficult for another side-chain to become protonated. Conversely, a negative J indicates positive cooperativity, meaning existing protonated sites make it easier for another side-chain to become protonated. Again, a calculation for D for Model 3 can be done following the same steps outlined above for Model 1.

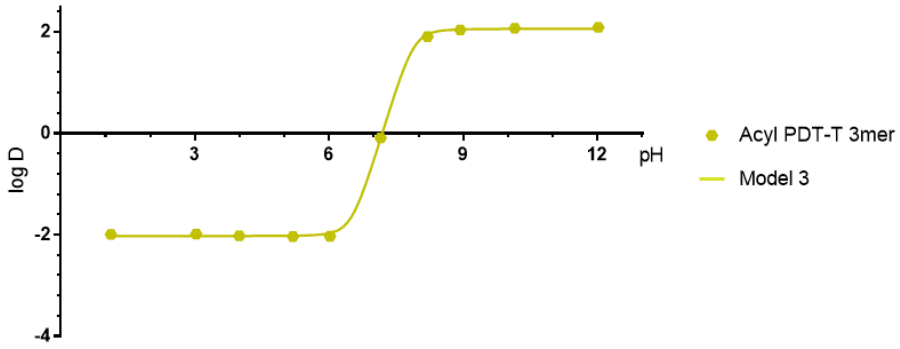
$$D = \frac{\sum_i P_{octanol,i}}{\sum_j P_{water,j}}$$

$$= \frac{(1 + e^{-\beta(\epsilon_{oct}-\mu)} + e^{-\beta(\epsilon_{oct}-\mu)} + e^{-\beta(\epsilon_{oct}-\mu)} + e^{-\beta(2\epsilon_{oct}-2\mu+J)} + e^{-\beta(2\epsilon_{oct}-2\mu+J)} + e^{-\beta(2\epsilon_{oct}-2\mu+J)} + e^{-\beta(3\epsilon_{oct}-3\mu+3J)})/Z}{(e^{-\beta\epsilon} + e^{-\beta(\epsilon_w-\mu+\epsilon)} + e^{-\beta(\epsilon_w-\mu+\epsilon)} + e^{-\beta(\epsilon_w-\mu+\epsilon)} + e^{-\beta(2\epsilon_w-2\mu+\epsilon+J)} + e^{-\beta(2\epsilon_w-2\mu+\epsilon+J)} + e^{-\beta(2\epsilon_w-2\mu+\epsilon+J)} + e^{-\beta(3\epsilon_w-3\mu+\epsilon+3J)})/Z}$$

$$= \frac{1 + 3e^{-\beta(\epsilon_{oct}-\mu)} + 3e^{-\beta(2\epsilon_{oct}-2\mu+J)} + e^{-\beta(3\epsilon_{oct}-3\mu+3J)}}{e^{-\beta\epsilon} + 3e^{-\beta(\epsilon_w-\mu+\epsilon)} + 3e^{-\beta(2\epsilon_w-2\mu+\epsilon+J)} + e^{-\beta(3\epsilon_w-3\mu+\epsilon+3J)}}$$

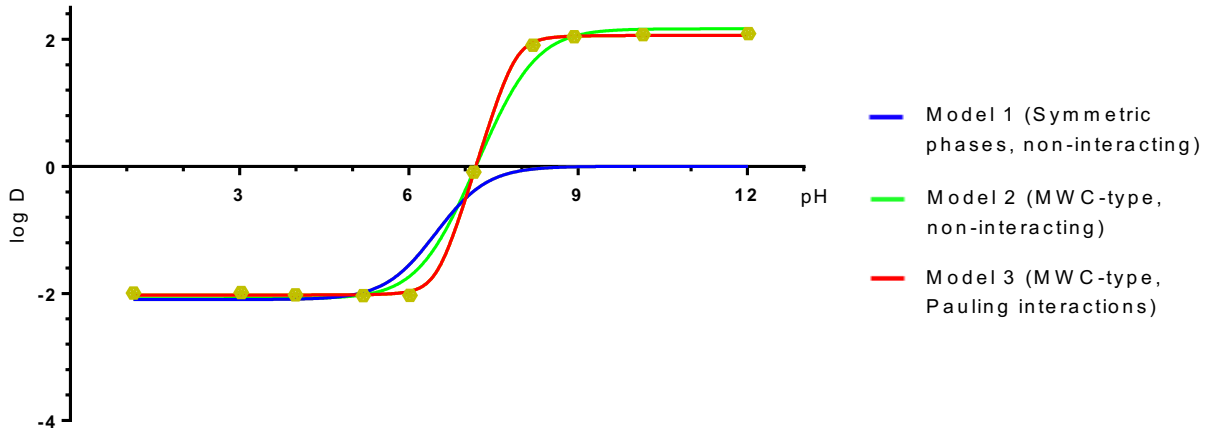
Following the same intermediate steps as Model 2, with this additional definition: $int \equiv \beta J$

$$\log D = \log\left[\frac{1 + 3e^{-oct-\frac{pH}{\log(e)}-\ln(c_0)} + 3e^{-2oct-int-2\frac{pH}{\log(e)}-2\ln(c_0)} + e^{-3oct-3int-3\frac{pH}{\log(e)}-3\ln(c_0)}}{e^{-diff} + 3e^{-water-diff-\frac{pH}{\log(e)}-\ln(c_0)} + 3e^{-2water-diff-int-2\frac{pH}{\log(e)}-2\ln(c_0)} + e^{-3water-diff-3int-3\frac{pH}{\log(e)}-3\ln(c_0)}}\right]$$



Revisiting the data set for the PDT-T 3-mer again, after fitting with Model 3, the inclusion of cooperativity allows the best fit to the experimental data yet. For this particular data set, this suggests

the protonation of a side chain is to some extent influenced by the protonation state of neighboring side chains.



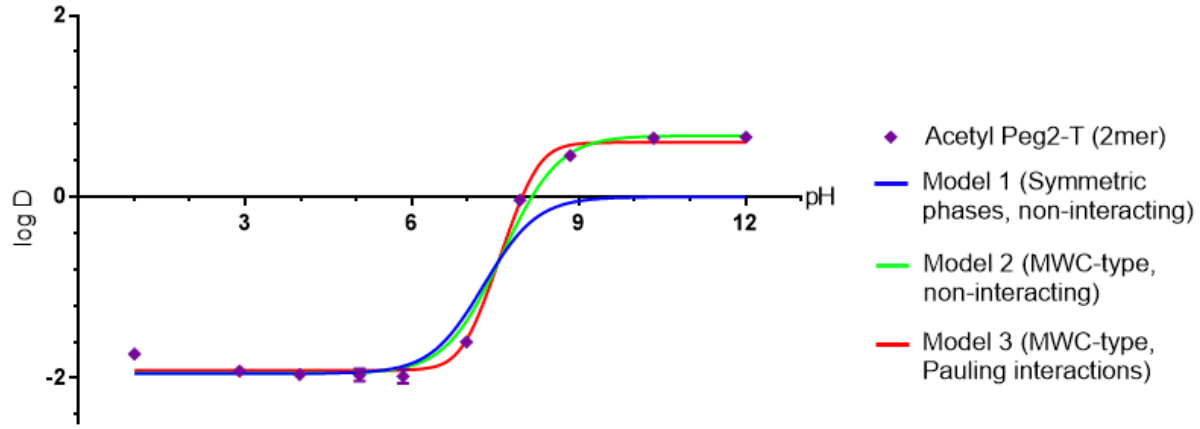
Model for 3 Sites	AICc
Model 1 (Symmetric phases, non-interacting)	18.14
Model 2 (MWC-type, non-interacting)	-67.86
Model 3 (MWC-type, Pauling interactions)	-118.6

While it was qualitatively easy to pick out the best model by visual inspection for this data set, I used the small-sample-size corrected version of the Akaike information criterion (AICc) to objectively select the best model (discussion of this criteria is done in the Comparing Models with AICc section of the Supporting Information). Comparing all 3 models, Model 3 best fits the experimental data from the 3 choices because a more negative AICc corresponds to a better model. This suggests that there must be an energy difference for the compound to be in the aqueous phase versus in the octanol phase and that there is some degree of interaction (cooperativity) between protonated groups on the compound.

It is important to note that Model 3 is also the most general model. In other words, in the appropriate limits, we can recover the other two models. Setting $J = 0$ in Model 3 recovers Model 2. Setting $J = 0, \epsilon = 0$ in Model 3 recovers Model 1. Thus, it shows that assumptions of independent sites and symmetric solvent phases cannot be made for the PDT-T 3-mer data set. The data suggests the J and ϵ parameters are needed to explain the data and not simply a means to overfit the data. For this particular compound, the models suggest there are indeed nanoscale intramolecular interactions between side-chains taking place, something that could not be revealed by only the bulk measurement.

Models for 2 Protonation Sites

One can repeat the same procedure used for 3 sites to derive the expressions for log D for 2 sites. The only real difference moving from a system with 3 protonation sites to a system of 2 protonation is in the enumeration of possible states the system can be in. The expressions for log D essentially take on the same form, just with different coefficients. The results for the 3 different models are shown below for the PEG2-T 2-mer data set.



Model for 2 Sites	AICc
Model 1 (Symmetric phases, non-interacting)	-31.39
Model 2 (MWC-type, non-interacting)	-72.16
Model 3 (MWC-type, Pauling interactions)	-84.34

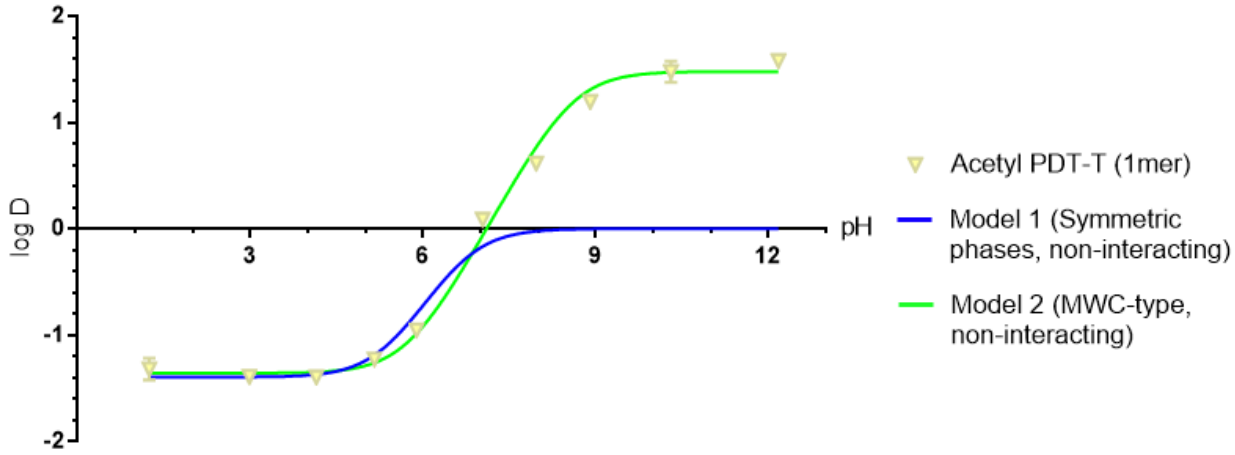
Again, I see the assumption of symmetric solvent phases results in the worst model. In this case, the inclusion of interaction terms resulted in the best model, suggesting for the PEG2-T 2-mer there should be cooperativity occurring between side-chains.

As a control, when I try to fit the 2-mer data with models made for 3 sites or 1 site, comparison of AICc shows that the fit will not be as models made for 2 sites. This is not the same as overfitting because changing the number protonation sites does not increase the number of parameters in the model; rather, it changes the weights and multiplicities of the states (see the Testing for Fitting Dependence on the Number of Protonation Sites section in the Supplementary Information). This indicates the Gibbs distribution should only correspond with accurate weights and multiplicities once you have defined the number of protonation states by properly enumerating the states.

Models for 1 Protonation Site

Finally, I play this game again for the limiting case of 1 protonation site. There still is the option of picking between symmetric solvent phases or a non-trivial energetic difference in

compound solvation energy between solvent phases. However, having only 1 protonation site excludes the possibility for interaction effects because there would not be another side-chain to physically interact with. I therefore limit this analysis to only Model 1 and Model 2 for the case of 1 protonation site. Shown below is the comparison of these models for the PDT-T 1-mer.



Model for 1 Site	AICc
Model 1 (Symmetric phases, non-interacting)	-0.798
Model 2 (MWC-type, non-interacting)	-83.82

As seen for the other 2 compounds, there must be a solvation energy difference to recapitulate the data for the PDT-T 1-mer.

While it is not a physically reasonable model, Model 3 was fit with data for the 1-mers. And in all cases, the AICc picks either Model 1 or Model 2 as better than Model 3, which is reassuring that the criteria penalizes models that use extra parameters to overfit the data (see Testing AICc on Overfitting of 1-mers section in the Supporting Information).

Analysis of Model Parameters in the oligoTEA Library

Armed with the log D expressions, I find the best model for each compound in the oligoTEA library from the 3 possible models according to the lowest AICc value. The AICc values for all models are shown below and further commented on in Comparing Models with AICc in the Supporting Information.

Oligomer	AICc Value Model 1 (2 parameters)	AICc Value Model 2 (3 parameters)	AICc Value Model 3 (4 parameters)	Notes on comparing AICc values between models
Acetyl-DTT-T (1 mer)	**-102.8**	-100.1		Model 1 has a probability ratio of 3.82 over Model 2 (79% to 21%)

Acetyl-DTT-T (2 mer)	-175	-193.1	**-206.8**	
Acetyl-DTT-T (3 mer)	-31.27	**-34.41**	-31.4	Model 2 has a probability ratio of 4.49 over Model 3 (82% to 18%). Model 2 has a probability ratio of 4.81 over Model 1 (83% to 17%)
Acetyl-PDT-T (1 mer)	-0.798	**-83.82**		
Acetyl-PDT-T (2 mer)	11.46	**-110.8**	**-109.7**	Model 2 has a probability ratio of 1.72 over Model 3 (63% to 37%)
Acetyl-PDT-T (3 mer)	18.14	-67.86	**-118.6**	
Acetyl-PEG2-T (1 mer)	-77.68	**-155**		
Acetyl-PEG2-T (2 mer)	-31.39	-72.16	**-84.34**	
Acetyl-PEG2-T (3 mer)	-12.07	**-36.14**	**-34.52**	Model 2 has probability ratio of 2.25 over Model 3 (69% to 31%)

The following table aggregates the best model for each compound and the corresponding value of the fitted parameters from that model. I look for correlations in the parameters and see if they can give us any insight on if interactions between side-chains are occurring. And if there are interactions, are they impacted by the molecular topology?

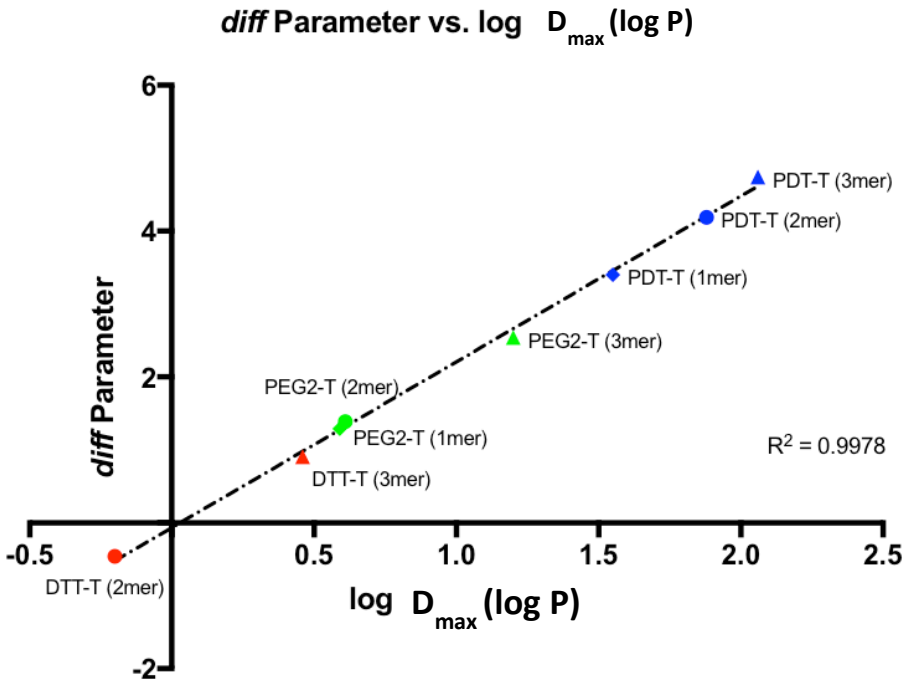
Compound	**Best-fitting Model**	**Parameters**			
		oct	*water*	*diff*	*int*
Acetyl-PDT-T (1 mer)	Model 2 (MWC-like, non-interacting)	2.868	-3.661	3.402	
Acetyl-PDT-T (2 mer)	Model 2 (MWC-like, non-interacting)	3.058	-2.11	4.19	
Acetyl-PDT-T (3 mer)	Model 3 (MWC-like, Pauling interactions)	3.516	0.382	4.741	-2.394
Acetyl-DTT-T (1 mer)	Model 1 (symmetric phases, non-interacting)	2.796	-2.791		
Acetyl-DTT-T (2 mer)	Model 3 (MWC-like, Pauling interactions)	-0.465	-2.597	-0.462	-1.928
Acetyl-DTT-T (3 mer)	Model 2 (MWC-like, non-interacting)	-0.599	-2.482	0.9045	
Acetyl-PEG2-T (1 mer)	Model 2 (MWC-like, non-interacting)	3.078	-2.195	1.288	
Acetyl-PEG2-T (2 mer)	Model 3 (MWC-like, Pauling interactions)	3.28	0.382	1.389	-6.363
Acetyl-PEG2-T (3 mer)	Model 2 (MWC-like, non-interacting)	-0.266	-3.489	2.539	

oct and water Parameters

There are a couple trends for the *oct* and *water* parameters within families of compounds, but there is no apparent global pattern. Recall the two parameters (defined as $oct \equiv \beta(\epsilon_{oct} - \mu_0)$ and $water \equiv \beta(\epsilon_w - \mu_0)$) indicate the energy associated with protonating a site while the compound is in the octanol or aqueous phase respectively. For *oct*, the value increases with mer length for the PDT family. For *water*, the value increases with mer length for the PDT and DTT family.

diff Parameter

It is quite clear the MWC-like models dominate the list, and the only case where a symmetric phase assumption seems to be applicable is the DTT 1mer. The *diff* parameter (defined as $diff \equiv \beta\epsilon$) reflects the difference in the energy of the compound to be solvated in the octanol phase versus in the aqueous phase. In hindsight, it makes logical sense that a difference in solvation energy is needed to help drive asymmetric partitioning of the compound into the two phases. From looking at the relative values of *diff* from fitting, we recapitulate the expected trend from knowing the relative hydrophobicities of the backbone dithiol comonomer. Indeed, comparing the values of *diff* with another parameter that reflects hydrophobicity, such as log D_{min} (log P), produces near perfect correlation.



Diff has the greatest values for the hydrophobic PDT family, the lowest values for the hydrophilic DTT family, and intermediate values for the PEG2 family. Across all families, *diff* increases with mer length. This suggests that in general, oligoTEAs with tertiary amine side-chains become more hydrophobic the longer their sequences. This is consistent with the idea that increasing the size of hydrophobic objects increases the free-energy cost of keeping them in a hydrophilic environment[51–53]. This parameter also shows that the backbone provides the dominate contribution to the compound's hydrophobicity because the *diff* parameter for the tertiary amine monomer is -0.29, which indicates it is slightly hydrophilic. In contrast, nearly all of the compounds in the library are hydrophobic.

int Parameter

For compounds where interactions between protonation sites are possible (i.e. the 2- and 3-mers), for half of those compounds the interacting Pauling model is the best fitting model. Based on the sign of the *int* parameter (defined as $int \equiv \beta J$), a positive value indicates a negative cooperativity (i.e. an additional protonation from the n-1 protonation state is more energetically costly than if the sites were independently protonated) and a negative value indicates positive cooperativity. The only discernable trend is that when there is cooperativity, it is of the positive form, where protonated side-chains make it easier for fellow side-chains to become protonated. There is even no continuity in the type of cooperativity among families. In the PEG2 and DTT families, going from the 2- to the 3- mers results in the loss of positive cooperativity. And in the PDT family, going from the 2- to the 3- mers results in gaining cooperativity. This implies cooperativity is a nonlinear function of mer length.

This can perhaps be explained by realizing that the effective shape of the molecule in solution is also a nonlinear function of the mer length. While a site might be favorably positioned for protonation in the 2-mer, the same site might see an entirely different local environment in the 3-mer. While the values of this parameter suggest prediction of interaction for other compounds would be very difficult, this generally does answer the question on the existence of electrostatic interactions between side-chains. The calculations suggest that interactions do exist for some of the compounds in this library.

A simple explanation as to why one should not be surprised that intramolecular interaction between side-chains is possible is through the heuristic of the Debye screening length. Since I used PBS 1x salt solution to make the aqueous layers, using a salt concentration of roughly 200mM, the

Debye length can be estimated to be ~0.7nm[50]. This is the effective maximum length to which a charge influences another charge in aqueous solution at that salt concentration. And since these oligoTEAs are known to be geometrically smaller than that length[54], then I should preliminary expect the possibility of electrostatic interaction. It is somewhat surprising then that not all of the multi-mers display some significant degree of interaction. However, the nature of that interaction will depend importantly on how the protonation sites are oriented with respect to each other in 3-D space. Studies have shown that relative orientation affects the interaction between hydrophobic and charged/polar surfaces[46]. While I cannot experimentally see how the sites are orientated with respect to each other, perhaps to some extent those effects might be in play here.

Conclusion

Motivated by the need for endosomal escape agents in applications such as siRNA drug delivery, this study sought to uncover some of the design principles related to controlling oligoTEA pK_a and hydrophobicity. Emphasis was placed on understanding possible interactions between side-chains and interactions between the side-chain and backbone. By designing a library of different backbone hydrophobicities and different mer lengths, I could begin investigating these interactions.

Using the shake flask method, I was able make bulk measurements of compound partitioning at microscopic resolution for an extended library of compounds. By repeating the method for aqueous solutions of different pH, one can trace out a log D curve for each compound. This methodology conveniently allows for measurement of pK_a and various parameters of hydrophobicity simultaneously from the log D curves.

Resulting analysis of the two parameters show some interesting trends that might be useful in design of an endosomal escape agent. The pK_a's of all compounds tested lie within the range of 6 to 8, which happens to be within the relevant window for pH maturation in the endosome. This suggests the tertiary amine could be a reasonable candidate for a side-chain group for that hypothesized agent.

The data suggests there is only a small degree of interaction between the side-chain and backbone. This is evidenced by the fact that among the 1-mers, the choice of backbone dithiol does not change the pK_a to any significant extent. The same observation is made for the 3-mers and to some extent the 2-mers. This is somewhat surprising given that the pK_a of a proxy for the tertiary amine subunit itself was measured to be about 2 units different from the tertiary amine when it is part of an oligoTEA allyl acrylamide monomer unit, which suggested the local environment of the backbone might affect the pK_a. This lack of interaction is shown also by library members with additional modifications. When the end cap of one of the oligomers was changed, there was no appreciable difference on its pK_a relative to the normal capped counterpart. And when 1.5 mers were made, which should have increased possible interactions between the side-chain and backbone over their integer counterparts, there was again no large change in pK_a. Perhaps then the dithiol comonomer is too far from the tertiary amine topologically to have a significant interaction.

Analysis of the pK_a and hydrophobicity parameters show that there is some correlation between parameter values of some composite oligoTEAs and the parameter values of the individual components that constitute them. The hydrophobicity of the backbone component seems to be crucial to the hydrophobicity of the overall construct. The relative hydrophobicity between choices

of backbone dithiol are preserved when looking at the log D_{min} and log P hydrophobicity values of the oligoTEA library. It is also seen the presence of additional backbone units (in the 1.5 mers) can significantly move these parameter values away from their integer counterparts. The *diff* parameter also suggests that the contribution of the backbone to hydrophobicity dominates over the side-chain's contribution because nearly all the library members are hydrophobic, whereas the *diff* value for the tertiary amine monomer shows it is slightly hydrophilic. Another interesting correlation was the span of log D as a function of size. The tertiary amine monomer had the smallest log D span. As one increased mer length, in general the span grew larger; this suggests increasing length allows one to increase the operating range of compound hydrophobicity. The existence of such correlations is useful from a design perspective in that they preserve the linear paradigm that the whole is the sum of its parts.

While I did not have direct evidence for interactions between side-chains, using theory I could infer the existence of such interactions for some of the members of the library. With a theoretical framework provided by equilibrium statistical mechanics, I made models that could calculate the equilibrium distributions for the partitioning of a compound with multiple protonation states into two phases. Using those distributions, I derived expressions for log D that could then be fitted with the experimental data. From fitting the parameters, for several of the multi-mers (PDT 3-mer, DTT 2-mer, and PEG2 2-mer), a non-trivial *int* parameter predicts there are intramolecular interactions between side-chains occurring. For these particular compounds, it suggests the nature of the interaction between side-chains is that of positive cooperativity, where existing protonated sites make it more energetically favorable for another side-chain to become protonated. One possible explanation could be an electrical double-layer effect where the protonated side-chain causes the formation of negative charge layer around the side-chain, which in term forms a positive charge layer around that. And perhaps by coincidence, the other side-chain(s) happens to be at a distance that falls within that positive layer, which makes protonation more likely.

Based on the compounds where interaction was inferred, it seems there is no discernable correlation between choice of backbone and interaction of the side-chains nor between mer length and interaction of the side-chains. While a rudimentary calculation of a Debye length hinted that charge-charge effects are possible for molecules of the size of these compounds[50], the general flexibility of oligoTEAs[54] makes prediction of these interactions from molecular topology alone seemingly difficult without some kind of knowledge of their arrangement in 3-D space and their

dynamics. In the following Outlook section, I propose some potential directions the design of charge-charge interactions for future oligoTEAs.

Outlook

Inference using the models allowed me to make the claim there are interactions between side-chains occurring in some of the multi-mers. However, here I am not able to measure such an interaction directly. It would be interesting to see if there are any experimental techniques that might be able to confirm the presence or characterize the nature of these electrostatic interactions. Techniques such as ESR allow one to determine the distance between two moieties on the same molecule, but it is not quite clear if the addition of probes needed for such experiments would necessarily disturb the natural interaction between side-chains.

While not pursued here, a logical next step beyond intramolecular interactions is studying the effects of intermolecular interactions. For instance, of interest would be characterizing cationic/zwitterion intermolecular interactions that may be typical between an endosomal escape agent and the lipids of the endosomal membrane. While it is difficult to imagine how that might be accomplished *in vivo*, techniques such as SPR might be able to shed some light on those interactions in an artificial setting.

For this library of compound, I restricted the side-chain to only being a tertiary amine. This was only done for convenience and simplicity, since it only has 1 ionization state. However, it would be interesting to study oligoTEAs with other side chains that have the possibility for carrying charge aside from the tertiary amine. It would be interesting to see if the models would reveal any side-chain/side-chain interactions between those groups.

References

1. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391,** 806–811 (1998).
2. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411,** 494–498 (2001).
3. Song, E. *et al.* RNA interference targeting Fas protects mice from fulminant hepatitis. *Nat. Med.* **9,** 347–351 (2003).
4. Davis, M. E. *et al.* Evidence of RNAi in humans from systemically administered siRNA via targeted nanoparticles. *Nature* **464,** 1067–1070 (2010).
5. The commercial tipping point. *Nat Biotech* **35,** 181–181 (2017).
6. Watts, J. K., Deleavey, G. F. & Damha, M. J. Chemically modified siRNA: tools and applications. *Drug Discov. Today* **13,** 842–855 (2008).
7. Adams, D. *et al.* Patisiran, an RNAi Therapeutic, for Hereditary Transthyretin Amyloidosis. *N. Engl. J. Med.* **379,** 11–21 (2018).
8. Commissioner, O. of the. Press Announcements - FDA approves first-of-its kind targeted RNA-based therapy to treat a rare disease. Available at: https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/UCM616518.htm. (Accessed: 14th September 2018)
9. Wittrup, A. & Lieberman, J. Knocking down disease: a progress report on siRNA therapeutics. *Nat. Rev. Genet.* **16,** 543–552 (2015).
10. Pecot, C. V., Calin, G. A., Coleman, R. L., Lopez-Berestein, G. & Sood, A. K. RNA interference in the clinic: challenges and future directions. *Nat. Rev. Cancer* **11,** 59–67 (2011).
11. Whitehead, K. A., Langer, R. & Anderson, D. G. Knocking down barriers: advances in siRNA delivery. *Nat. Rev. Drug Discov.* **8,** 129–138 (2009).

12. Wang, T. *et al.* Challenges and opportunities for siRNA-based cancer treatment. *Cancer Lett.* **387,** 77–83 (2017).
13. Garber, K. Worth the RISC? *Nat. Biotechnol.* **35,** 198–202 (2017).
14. Juliano, R. L. The delivery of therapeutic oligonucleotides. *Nucleic Acids Res.* **44,** 6518–6548 (2016).
15. D'Souza, A. A. & Devarajan, P. V. Asialoglycoprotein receptor mediated hepatocyte targeting — Strategies and applications. *J. Controlled Release* **203,** 126–139 (2015).
16. Cummings, R. D. & McEver, R. P. C-type Lectins. in *Essentials of Glycobiology* (eds. Varki, A. et al.) (Cold Spring Harbor Laboratory Press, 2009).
17. Schwartz, A. L., Fridovich, S. E. & Lodish, H. F. Kinetics of internalization and recycling of the asialoglycoprotein receptor in a hepatoma cell line. *J. Biol. Chem.* **257,** 4230–4237 (1982).
18. Wittrup, A. *et al.* Visualizing lipid-formulated siRNA release from endosomes and target gene knockdown. *Nat. Biotechnol.* **33,** 870–876 (2015).
19. Saleh, M.-C. *et al.* The endocytic pathway mediates cell entry of dsRNA to induce RNAi silencing. *Nat. Cell Biol.* **8,** 793–802 (2006).
20. Doherty, G. J. & McMahon, H. T. Mechanisms of Endocytosis. *Annu. Rev. Biochem.* **78,** 857–902 (2009).
21. Liang, W. & W. Lam, J. K. Endosomal Escape Pathways for Non-Viral Nucleic Acid Delivery Systems. in *Molecular Regulation of Endocytosis* (ed. Ceresa, B.) (InTech, 2012).
22. Varkouhi, A. K., Scholte, M., Storm, G. & Haisma, H. J. Endosomal escape pathways for delivery of biologicals. *J. Controlled Release* **151,** 220–228 (2011).
23. Dominska, M. & Dykxhoorn, D. M. Breaking down the barriers: siRNA delivery and endosome escape. *J Cell Sci* **123,** 1183–1189 (2010).
24. Dowdy, S. F. Overcoming cellular barriers for RNA therapeutics. *Nat. Biotechnol.* **35,** 222–229 (2017).

25. Stewart, M. P., Langer, R. & Jensen, K. F. Intracellular Delivery by Membrane Disruption: Mechanisms, Strategies, and Concepts. *Chem. Rev.* **118,** 7409–7531 (2018).
26. Skehel, J. J. & Wiley, D. C. Receptor Binding and Membrane Fusion in Virus Entry: The Influenza Hemagglutinin. *Annu. Rev. Biochem.* **69,** 531–569 (2000).
27. Wiley, D. C. & Skehel, J. J. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annu. Rev. Biochem.* **56,** 365–394 (1987).
28. Jackson, A. L. & Linsley, P. S. Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nat. Rev. Drug Discov.* **9,** 57–67 (2010).
29. Judge, A. D. *et al.* Sequence-dependent stimulation of the mammalian innate immune response by synthetic siRNA. *Nat. Biotechnol.* **23,** 457–462 (2005).
30. Yin, H. *et al.* Non-viral vectors for gene-based therapy. *Nat. Rev. Genet.* **15,** 541–555 (2014).
31. Rozema, D. B. *et al.* Dynamic PolyConjugates for targeted in vivo delivery of siRNA to hepatocytes. *Proc. Natl. Acad. Sci.* **104,** 12982–12987 (2007).
32. Wong, S. C. *et al.* Co-Injection of a Targeted, Reversibly Masked Endosomolytic Polymer Dramatically Improves the Efficacy of Cholesterol-Conjugated Small Interfering RNAs In Vivo. *Nucleic Acid Ther.* **22,** 380–390 (2012).
33. Porel, M. & Alabi, C. A. Sequence-Defined Polymers via Orthogonal Allyl Acrylamide Building Blocks. *J. Am. Chem. Soc.* **136,** 13162–13165 (2014).
34. Porel, M., Thornlow, D. N., Phan, N. N. & Alabi, C. A. Sequence-defined bioactive macrocycles via an acid-catalysed cascade reaction. *Nat. Chem.* **8,** 590–596 (2016).
35. Porel, M., Thornlow, D. N., Artim, C. M. & Alabi, C. A. Sequence-Defined Backbone Modifications Regulate Antibacterial Activity of OligoTEAs. *ACS Chem. Biol.* **12,** 715–723 (2017).
36. Sorkin, M. R., Walker, J. A., Brown, J. S. & Alabi, C. A. Versatile Platform for the Synthesis of Orthogonally Cleavable Heteromultifunctional Cross-Linkers. *Bioconjug. Chem.* **28,** 907–912 (2017).

37. Tibbitt, M. W., Dahlman, J. E. & Langer, R. Emerging Frontiers in Drug Delivery. *J. Am. Chem. Soc.* **138,** 704–717 (2016).

38. Gori, J. L. *et al.* Delivery and Specificity of CRISPR-Cas9 Genome Editing Technologies for Human Gene Therapy. *Hum. Gene Ther.* **26,** 443–451 (2015).

39. Geisow, M. J. & Evans, W. H. pH in the endosome. *Exp. Cell Res.* **150,** 36–46 (1984).

40. Alabi, C. A. *et al.* Multiparametric approach for the evaluation of lipid nanoparticles for siRNA delivery. *Proc. Natl. Acad. Sci.* **110,** 12881–12886 (2013).

41. Zhang, J., Fan, H., Levorse, D. A. & Crocker, L. S. Ionization Behavior of Amino Lipids for siRNA Delivery: Determination of Ionization Constants, SAR, and the Impact of Lipid pKa on Cationic Lipid−Biomembrane Interactions. *Langmuir* **27,** 1907–1914 (2011).

42. Jayaraman, M. *et al.* Maximizing the Potency of siRNA Lipid Nanoparticles for Hepatic Gene Silencing In Vivo. *Angew. Chem. Int. Ed.* **51,** 8529–8533 (2012).

43. Waring, M. J. Lipophilicity in drug discovery. *Expert Opin. Drug Discov.* **5,** 235–248 (2010).

44. Pratt, L. R. & Pohorille, A. Hydrophobic Effects and Modeling of Biophysical Aqueous Solution Interfaces. *Chem. Rev.* **102,** 2671–2692 (2002).

45. Busscher, H. J., Norde, W. & Mei, H. C. van der. Specific Molecular Recognition and Nonspecific Contributions to Bacterial Interaction Forces. *Appl. Environ. Microbiol.* **74,** 2559–2564 (2008).

46. Ma, C. D., Wang, C., Acevedo-Vélez, C., Gellman, S. H. & Abbott, N. L. Modulation of hydrophobic interactions by proximally immobilized ions. *Nat. Lond.* **517,** 347-350M (2015).

47. Bolt, H. L. *et al.* Log D versus HPLC derived hydrophobicity: The development of predictive tools to aid in the rational design of bioactive peptoids. *Pept. Sci.* **108,** n/a-n/a (2017).

48. Perrin, D. D., Dempsey, B. & Serjeant, E. P. *pK a Prediction for Organic Acids and Bases*. (Springer Netherlands, 1981). doi:10.1007/978-94-009-5883-8

49. Acevedo-Vélez, C., Andre, G., Dufrêne, Y. F., Gellman, S. H. & Abbott, N. L. Single-Molecule Force Spectroscopy of β-Peptides That Display Well-Defined Three-Dimensional Chemical Patterns. *J. Am. Chem. Soc.* **133,** 3981–3988 (2011).
50. Phillips, R., Kondev, J., Theriot, J. & Garcia, H. *Physical Biology of the Cell, Second Edition*. (Garland Science, 2012).
51. Ball, P. Biophysics: More than a bystander. *Nature* **478,** 467–468 (2011).
52. Chandler, D. Hydrophobicity: Two faces of water. *Nature* **417,** 491–491 (2002).
53. Chandler, D. Interfaces and the driving force of hydrophobic assembly. *Nat. Lond.* **437,** 640–7 (2005).
54. Brown, J. S. *et al.* Synthesis and Solution-Phase Characterization of Sulfonated Oligothioetheramides. *Macromolecules* (2017). doi:10.1021/acs.macromol.7b01915
55. Hitzel, L., Watt, A. P. & Locker, K. L. An increased throughput method for the determination of partition coefficients. *Pharm. Res.* **17,** 1389–1395 (2000).
56. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19,** 716–723 (1974).
57. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. (Springer Science & Business Media, 2007).

Supplementary Information

Material and Methods

i) N,N-dimethylethylamine monomer (tertiary amine monomer) synthesis

Synthesis of the tertiary amine monomer is done in two main steps (acylation and alkylation) outlined as follows:

Acylation Step

Start with clean and dry glassware, stir bars, and starting materials. Into a round bottom flask, add 1g of N,N-dimethylethylenediamine. In a separate vial, dilute 1.2 molar equivalents of acryl chloride into 5 mL of dichloromethane (DCM) and set aside. Add 1.25 molar equivalents of trimethylamine (TEA) into the round bottom flask and enough DCM to make the final reaction mixture 0.15M (including the volume of the diluted acryl chloride). Seal the flask with a scepter and stir for 10 minutes. Cool the mixture down to ~0°C on ice and let sit for 10-15 minutes. Add the diluted acryl chloride to the flask dropwise for 1 hour at ~0°C. Then, stir the mixture on ice for 1 hour. After, stir the mixture at room temperature for 1 hour. Quench the mixture with 10N NaOH. Extract the product into the organic layer using DCM. Use TLC to monitor extraction of product into organic layer. Use sodium sulfate to dry any residual water that may be left in the organic layer. Transfer the organic mixture to a new flask, and dry off the product.

Alkylation Step

Start with clean and dry glassware, stir bars, and starting materials. In a vial, solubilize 1.2 molar equivalents of allyl bromide in 1 mL of dry dimethylformamide (DMF) and set aside. In a round bottom flask, solubilize 1.5 molar equivalents of sodium hydride (NaH) in 10 mL of DMF. In a separate vial, solubilize the product of the acylation step in enough to DMF to make a 0.2M total reaction mixture (including the volumes of solubilized NaH and solubilized allyl bromide). Add the solubilized acylation product to the flask containing NaH. Stir the flask for 10 minutes at room temperature. Then stir at ~0°C on ice, allowing the mixture to equilibrate for 15 minutes. Using a syringe, add the solubilized allyl bromide to the flask dropwise for ~45 minutes while keeping the flask on ice. Stir the flask on ice for 1 hour. Then, quench the mixture with a small volume of 1M HCl. Dry the contents of the flask, removing as much of the DMF as possible. Basify the mixture with 10N NaOH (making sure to check the pH of the aqueous becomes ≥ 14). Add MilliQ water to take up excess salts (checking that the water layer remains pH ≥ 14) Extract the product with DCM.

Use TLC to monitor extraction of product into organic layer. Use sodium sulfate to dry any residual water that may be left in the organic layer. Dry the product. Purify the product using ISCO purification ((optional hexanes wash) dichloromethane/ultra system) on a silica or basic alumina column.

ii) Oligomer synthesis

Synthesis of oligoTEAs follows a few main steps, as shown below. The methodology is adapted from one set forth by Porel et al[33]:

Fluorous solid-phase extraction (FSPE)

Dissolve the fluorous organic mixture to be separated with a small volume (1 equivalent) of methanol. Load that mixture onto a pre-packed fluorous solid phase extraction (FSPE) cartridge that has 1 equivalent of water on top of the column. Add 3 more equivalents of methanol to the top of the column and mix. Use a fluorophobic wash (80% methanol, 20% water) to elute all the non-fluorous molecules, leaving the fluorous molecules retained on the fluorous silica gel. Use 2-5mL of fluorophobic wash per 10mg of fluorous-tagged product. Then use a fluorophilic wash (100% methanol) to elute the fluorous molecules from the fluorous stationary phase. Use 1-2mL of methanol per 10mg of fluorous-tagged product.

Fluorous Boc-protected allyl amine

Into 10 mL THF, solubilize 100 mg (~0.147mmol) of 2-[2-(1H,1H,2H,2H-Perfluoro-9-methyldecyl)isopropoxycarbonyloxyimino]-2-phenylacetonitrile (fluorous BOC-ON), 1.2 molar equivalents of allyl amine, and 2 molar equivalents of TEA. Stir the reaction mixture at room temperature for at least 3 hours. Purify the product mixture using the FSPE process mentioned above in a 2g FSPE cartridge (using 20 mL of wash and eluting with 15 mL of methanol). Then dry off the product.

Thiol-ene reaction

In enough methanol to make a 100mM reaction mixture, add 1 molar equivalent (either fluorous Boc-protected allyl amine or a Michael addition product), 5 equivalents of dithiol, and 5 mol% (relative to dithiol) of 2,2-dimethoxy-2-phenylacetophenone (DMPA). Irradiate the reaction mixture for 270 consecutive seconds at 20 mW/cm^2. Stir the reaction vial to mix the contents.

Irradiate for an additional 90 seconds at mW/cm$^{2.}$ Purify the product mixture using the FSPE process mentioned above. Then dry off the product.

Thiol-Michael addition

In enough methanol to make a 100mM reaction mixture, add 1 molar equivalent of fluorous-thiol, 2 equivalents of tertiary amine monomer, and 5 mol% (relative to monomer) of dimethyl phenyl phosphine (Me_2PhP). Stir the mixture at room temperature for ~1 hour. The temperature and reaction time can be modified by knowing the kinetic data for the monomer on a specific thiol. Check for reaction completion by performing the DTDP assay. Purify the product mixture using the FSPE process mentioned above. Then dry off the product.

Dithio-dipyridyl (DTDP) assay

For each test, add 500 uL of a solution of 0.1% TEA in dimethyl sulfoxide (DMSO) to 8.33 μL of 12 mM 2,2'-dipyridyldisulfide in DMSO. Add 0.5 μL of 0.1M Michael addition reaction mixture and allow to sit for 5 minutes. Add 16.67 μL acetic acid and allow the mixture to react for 5 minutes. Make a blank using the same components but with 0.5 μL methanol instead of 0.5 μL Michael reaction mixture. Pipette volumes from each assay in triplicate. Obtain absorbance values of the blank and test mixtures by scanning from 330-449 nm using an absorbance plate reader. Look for the reactant peak to decrease as a function of reaction time. If the curves of absorbance versus wavelength for the blank and test mixtures overlap, the reaction is deemed to be complete.

Acetylation reaction

Cleaved oligoTEA with primary amine (1.0 equiv), acetic anhydride (2.0 equiv.), and TEA (2.0 equiv.) in DCM (0.5 M) were added into a 4mL vial. The mixture was periodically shaken at room temperature for 2 hours. The reaction mixture was purified using HPLC.

iii) log D shake-flask method to obtain distribution coefficients and pK_a

Experimental log D values were obtained using an adapted Eppendorf shake-flask method[55]. Take a known mass of analyte (quantified via qNMR using a benzene standard) and dissolve in 600uL of Milli-Q water (Solution A). Make 10 different pH solutions using PBS 1x buffer and titrating with HCl and NaOH. In a 2mL screw-top vial, add 50uL of Solution A; then add 550 uL of appropriate pH solution and 600uL of octan-1-ol. Repeat for a total of 10 different vials for each pH

solution (Test Solutions). For 2 other 2mL vials, add 50uL of Solution A to each. To one add 550uL of Milli-Q water; to the other, dry off the solvent and resuspend in 600uL of octan-1-ol (Standards 1). These are standards for each phase with concentrations on the order of the 10 different pH vials. From each standard, take 100uL and mix with 900uL of Milli-Q water/octan-1-ol for a 10x serial dilution (Standards 2). Perform an additional 10x serial dilution (Standards 3). Vortex the Test Solutions for 1 minute and sonicate for an additional 15 minutes. This will allow the analyte to equilibrate between the 2 phases, so excess mixing is allowed. If either phase appears cloudy, centrifuge (<3000 rpm) the vials in multiples of 1 minute until both phases appear clear; a cloudy phase indicates water/octanol emulsions, which would distort phase quantifications later. Pipette 500uL from each phase into a 2mL screw-top vial for detection (Detection Solutions).

The analyte in Detection Solutions, Standards 1/2/3, pure Milli-Q water, pure PBS 1x buffer, pure octan-1-ol is detected in 5uL injections using LC-MS. SIM for the M+H mass (or half mass if it presents a stronger signal) to 1 or 2 decimal places for improved sensitivity.

From the SIM data, obtain the area of analyte and tabulate for all samples. From the areas of the Milli-Q water and octan-1-ol standards, subtract the areas of the pure Milli-Q and pure octan-1-ol for a baseline correction. Using the known concentrations of the standards and areas create a standard curve for each phase. From the areas of Detection Solutions, subtract the areas of the pure PBS 1x buffer and pure octan-1-ol respectively for a baseline correction. Using the standard curves and baseline corrected areas, calculate the concentration for each phase at each pH. The ratio of the analyte detected in the octan-1-ol phase to the aqueous phase at each pH is the D value. Using statistical software (PRISM), plot log D vs. pH and fit a sigmoidal equation to the data. The IC50 value is the pK_a of the compound.

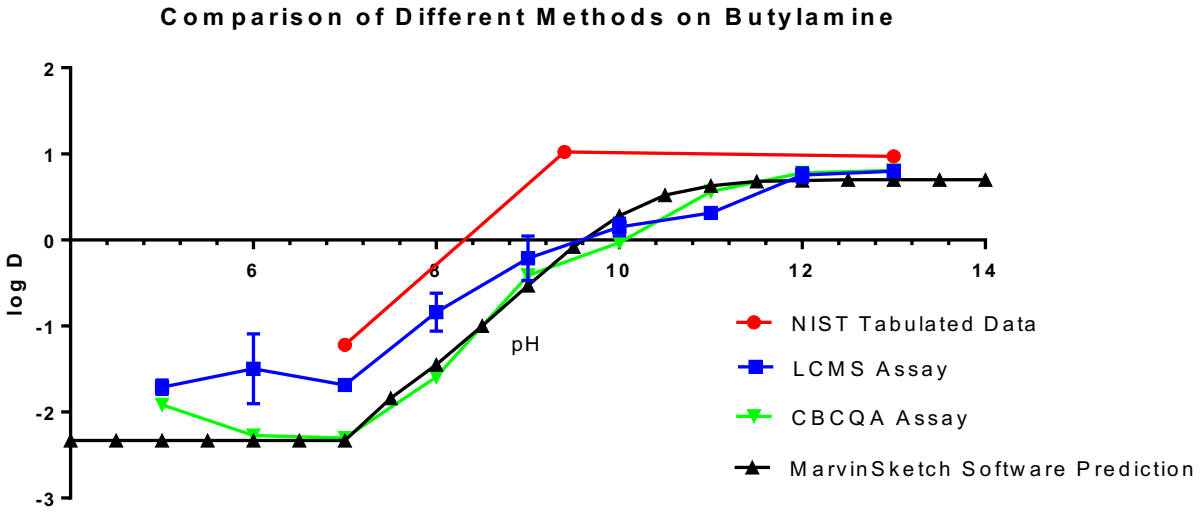
Control Experiments

LC-MS quantification

Usage of the shake-flask method in the literature has not typically relied on quantification with high resolution methods such as LC-MS. Since we are limited on the scale of analyte available, we needed a method that could quantify down to a resolution of 100 ng/mL. The rough estimation of the resolution of the LC-MS was done by quantifying serial dilutions of analyte used for standard curves in log D experiments. Based on known concentrations of the analyte, in SIM mode, the LC-MS can indeed detect down to the order of 100ng/mL for compounds with tertiary amines.

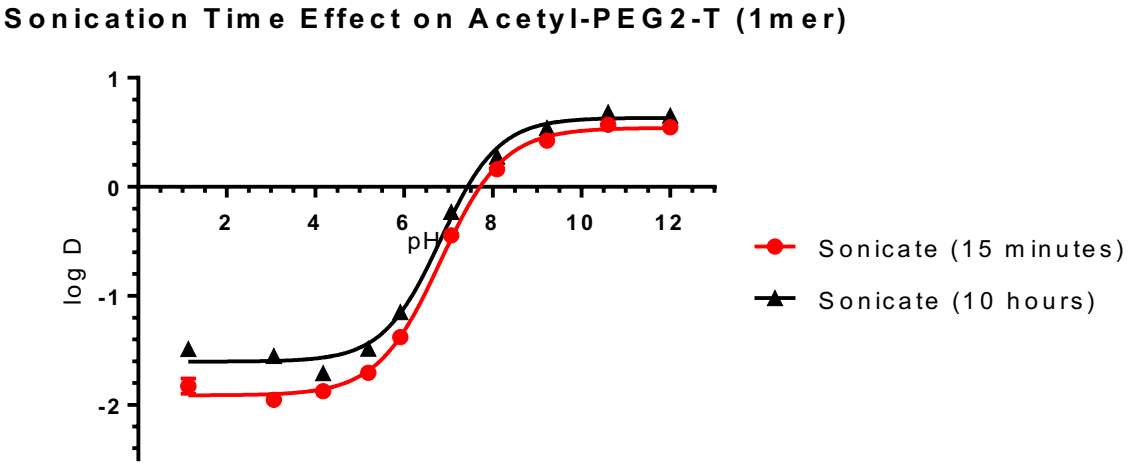
LC-MS benchmarking

To test the viability of the LC-MS as a quantification method, initial experiments on a compound with a known log-D profile was performed, 1-aminobutane. Relative agreement with literature values[48] and other methods suggest LC-MS can be used for future experiments.



Sonication time

The protocol listed above mentions to sonicate the Test Solutions for 15 minutes. One might wonder if that is an adequate time to allow the compound to equilibrate. Sonication for 15 minutes was compared against sonication for 10 hours. The results showed a small quantitative but not qualitative difference, suggesting the 15 minute time is adequate enough.

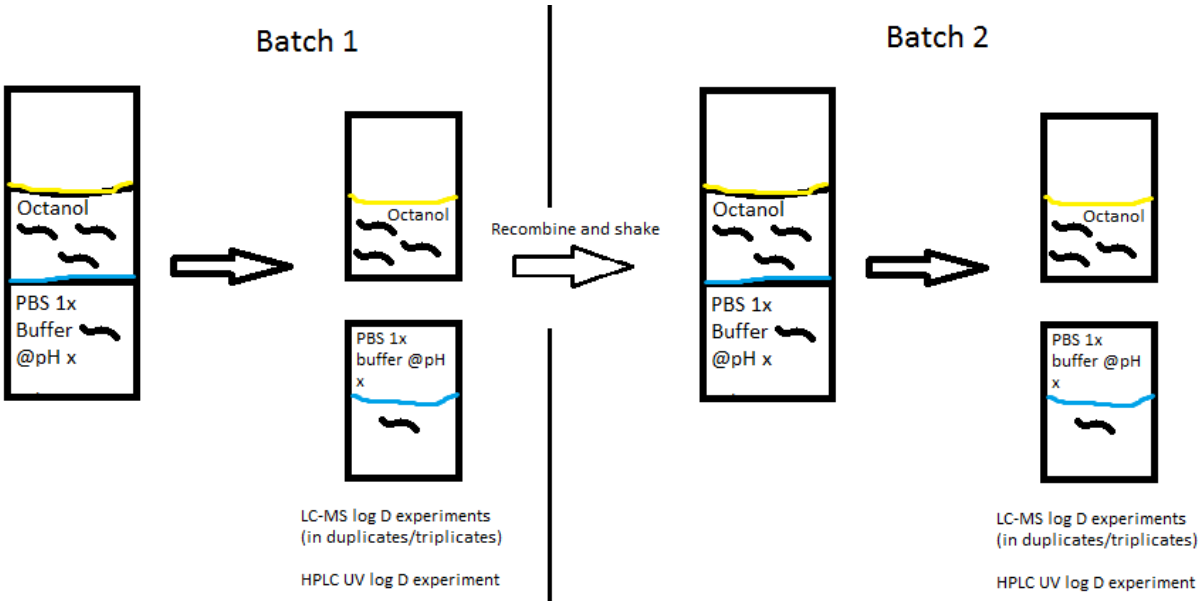


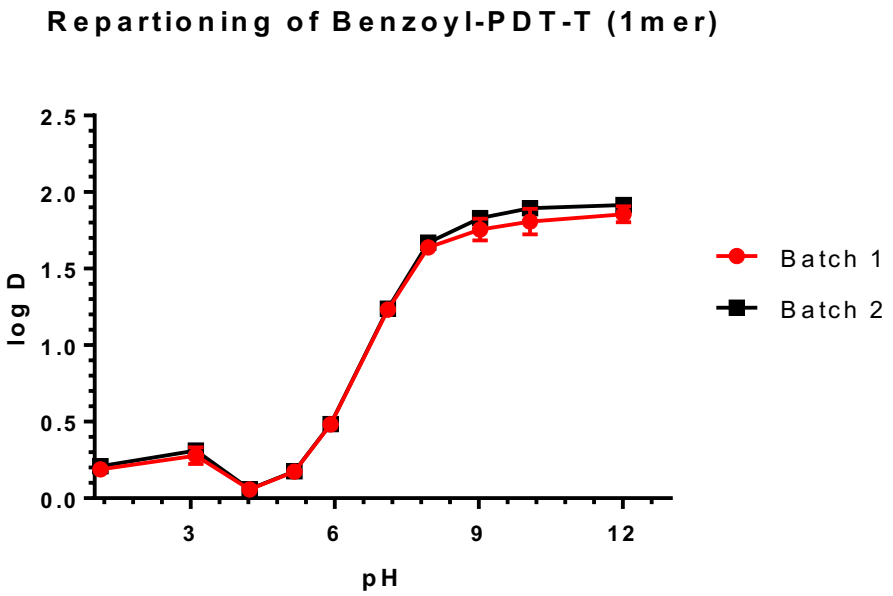
Centrifugation speed

The protocol listed above mentions to centrifuge the Test Solutions to de-mix the octanol and aqueous phases. One concern one might have is the larger mass of the analyte compounds may cause them to artificially accumulate in the denser aqueous phase during centrifugation. To prove this does not happen to an equal volume mixture of PBS 1x / octanol for the operating speeds of the centrifuge used (<3000 rpm), a fluorescent dye (FITC) was allowed to partition between the aqueous and octanol phase. Measuring the partition at various centrifugation speeds showed no change in the partition (data not shown).

Partitioning occurs at equilibrium

The utility of the partitioning relies on the fact that it should be an equilibrium measurement. One way to prove that the measurements are made at equilibrium is by doing the measurement as mentioned in the protocols above. Then take the octanol and aqueous phase samples and re-mix them. Then proceed through the protocol again with this re-mixed Test Solution. If the 2nd measurement matches the initial measurement, that indicates the partition achieved the same thermal equilibrium state.

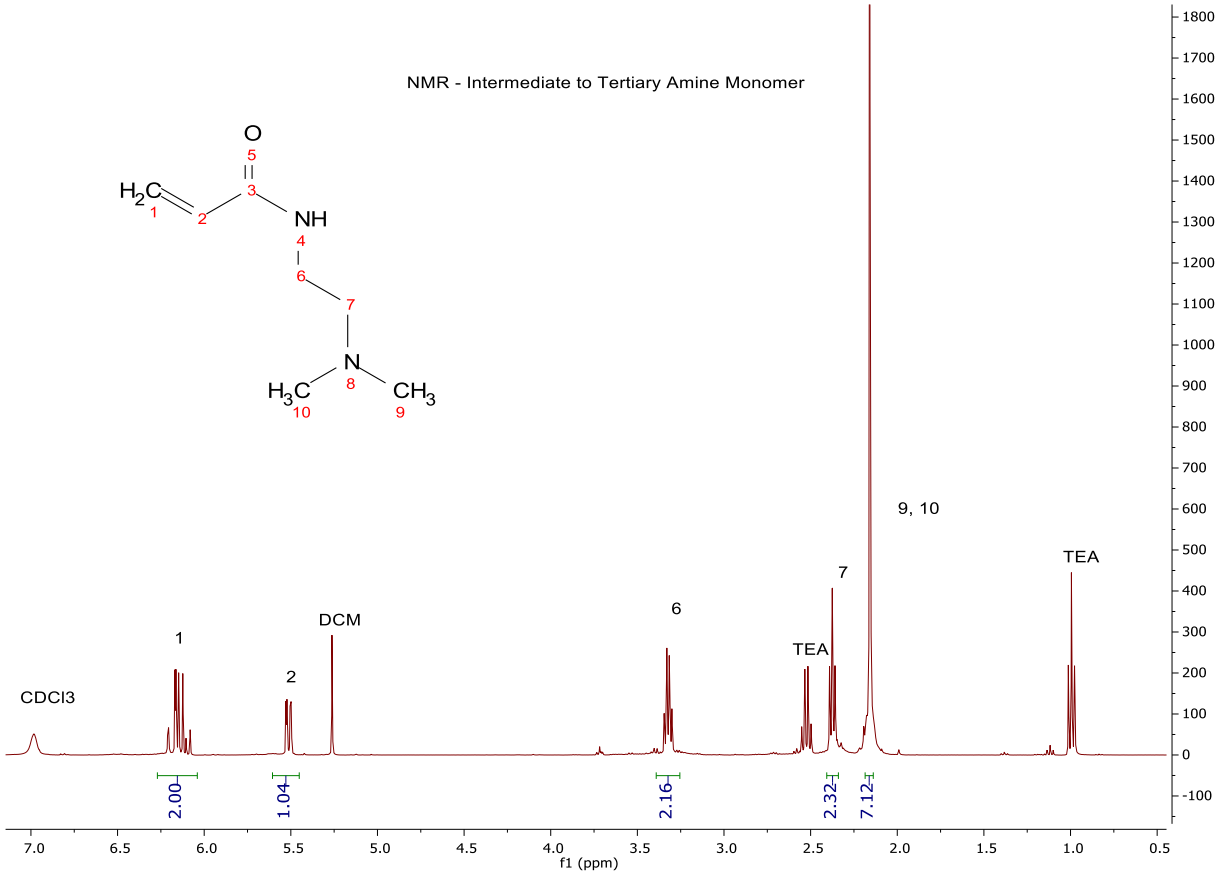


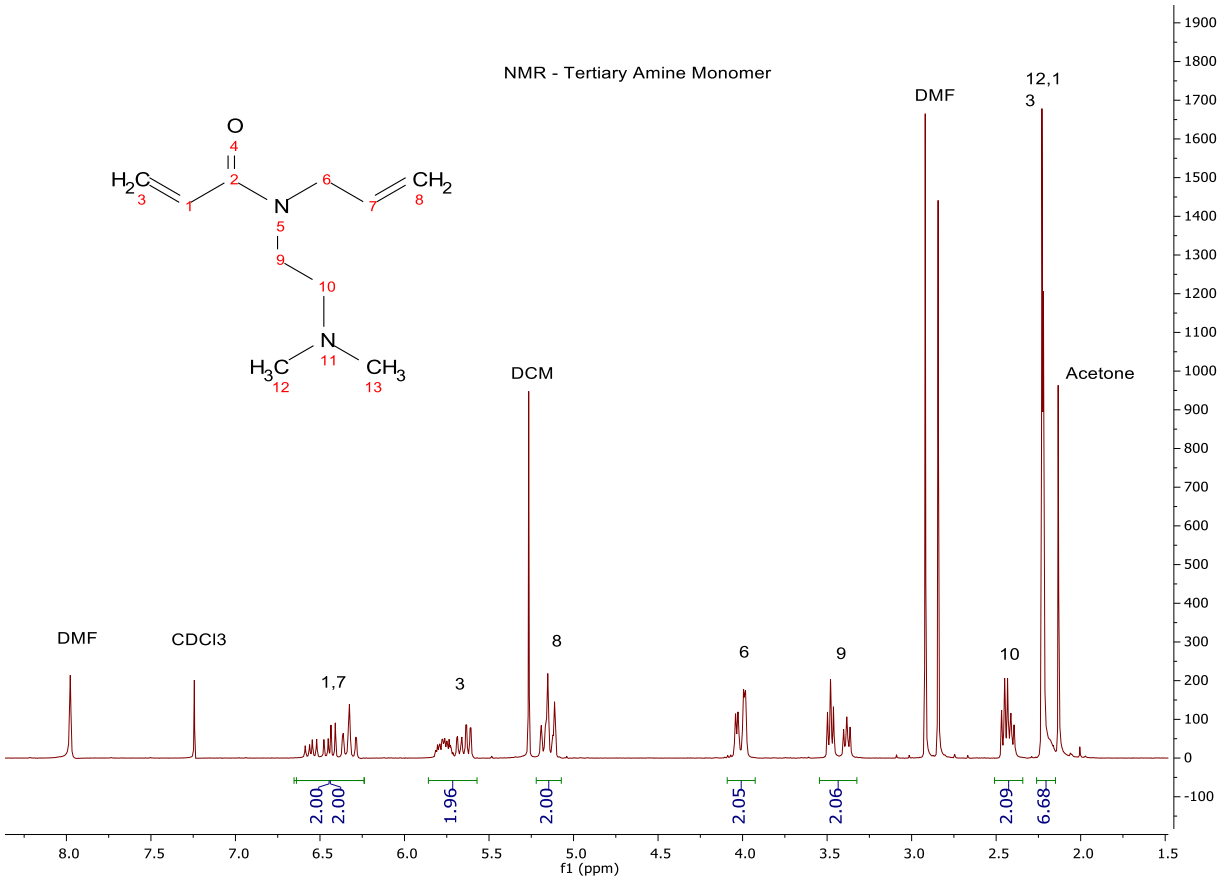


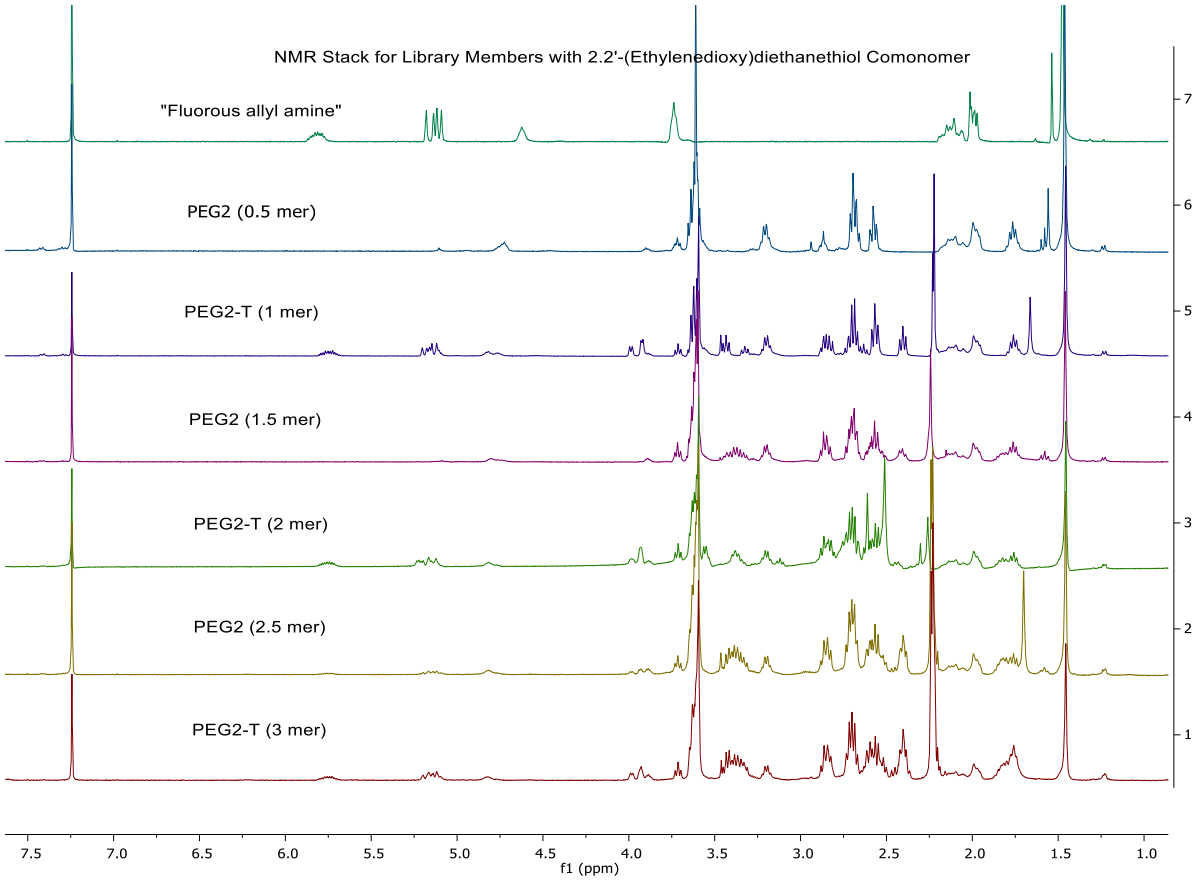
Nearly identical results for both batches indicate the partitioning achieves thermal equilibrium in both cases before the measurements are done.

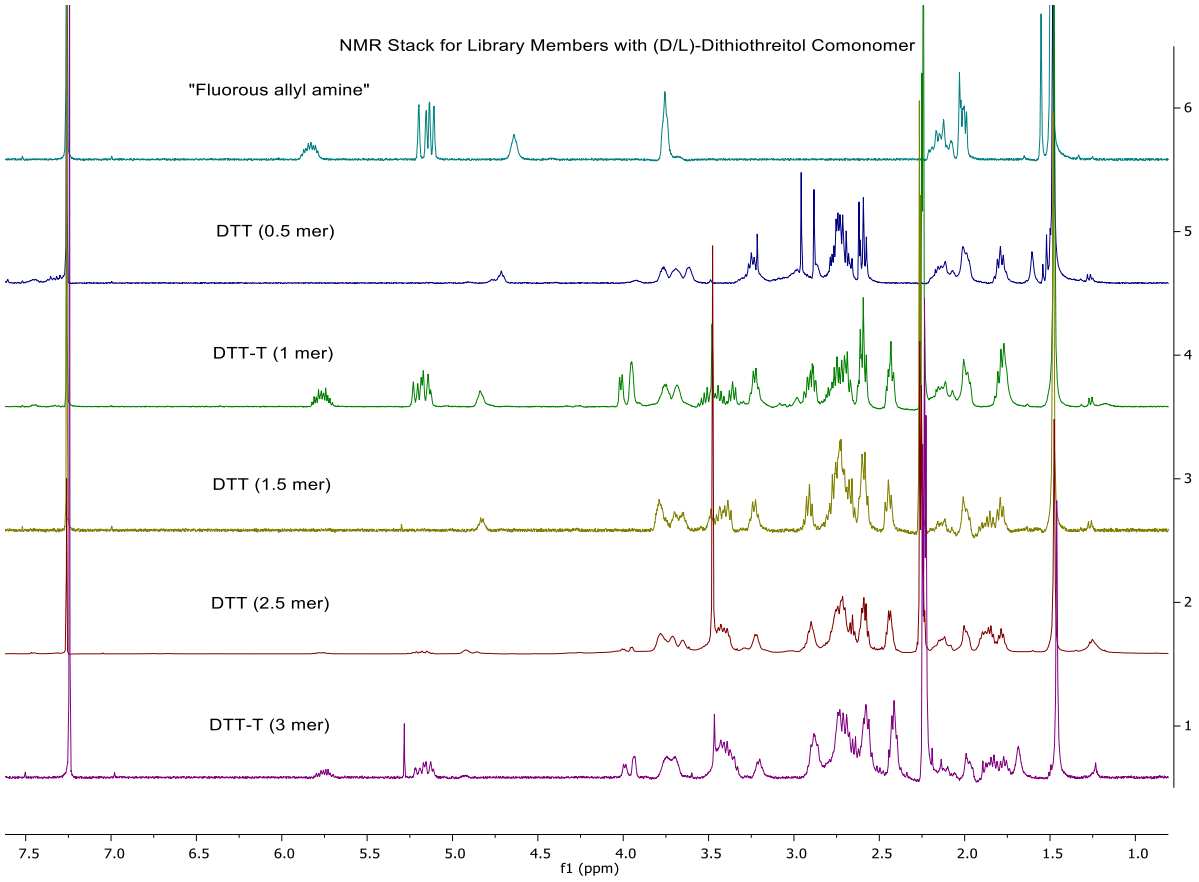
Compound Verification

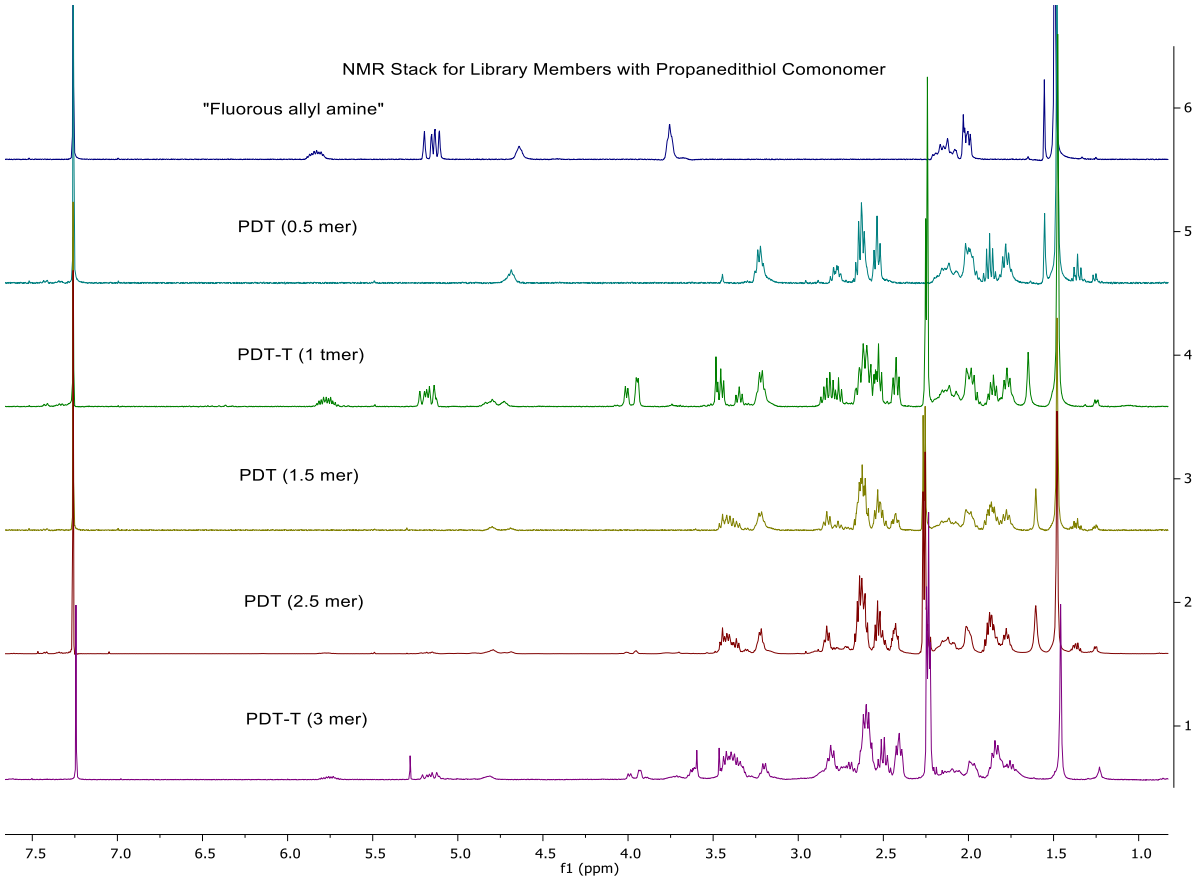
NMR



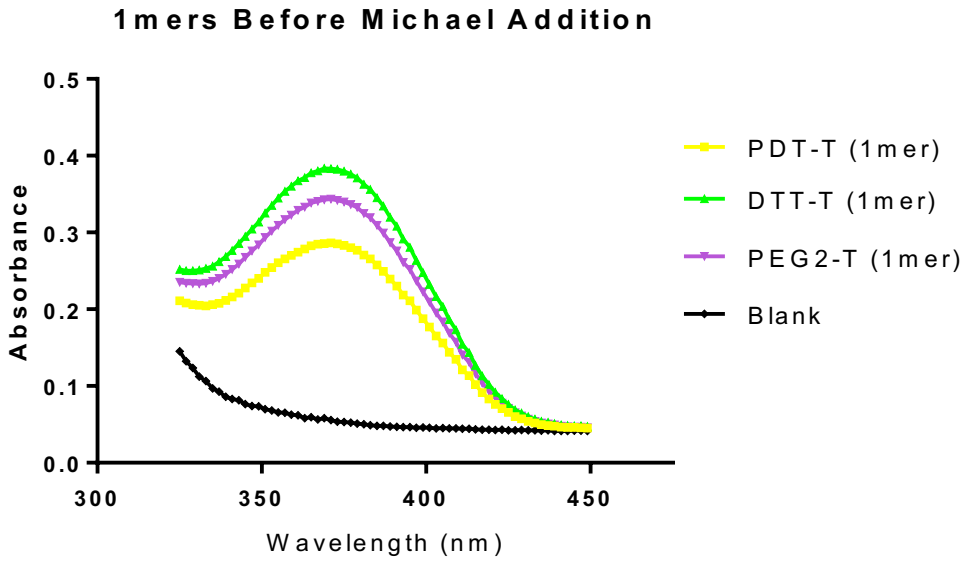

NMR - Tertiary Amine Monomer
DMF
CDCl3
1,7
3
DCM
8
6
9
DMF
10
12,1
3
Acetone
2.00
2.00
1.96
2.00
2.05
2.06
2.09
6.68
f1 (ppm)

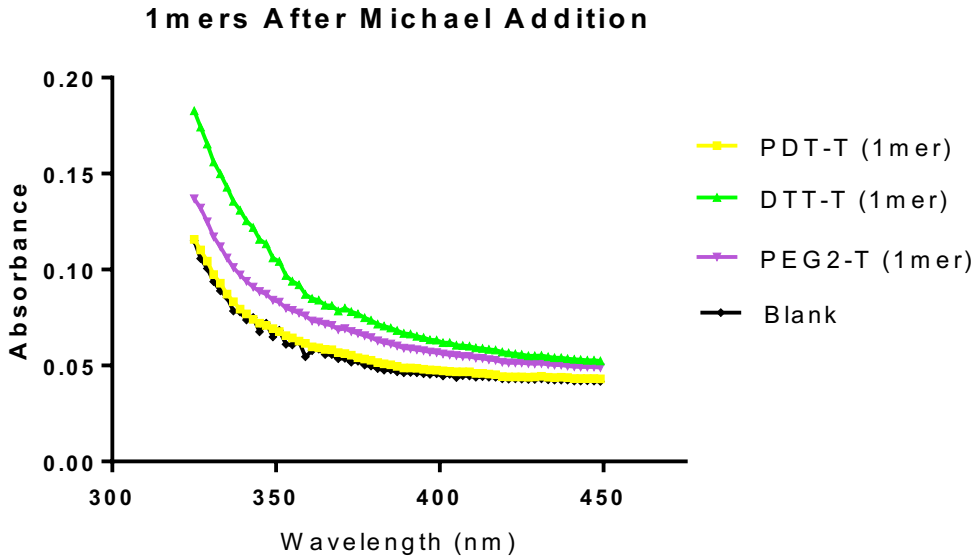

NMR Stack for Library Members with 2.2'-(Ethylenedioxy)diethanethiol Comonomer
"Fluorous allyl amine"
PEG2 (0.5 mer)
PEG2-T (1 mer)
PEG2 (1.5 mer)
PEG2-T (2 mer)
PEG2 (2.5 mer)
PEG2-T (3 mer)
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
f1 (ppm)

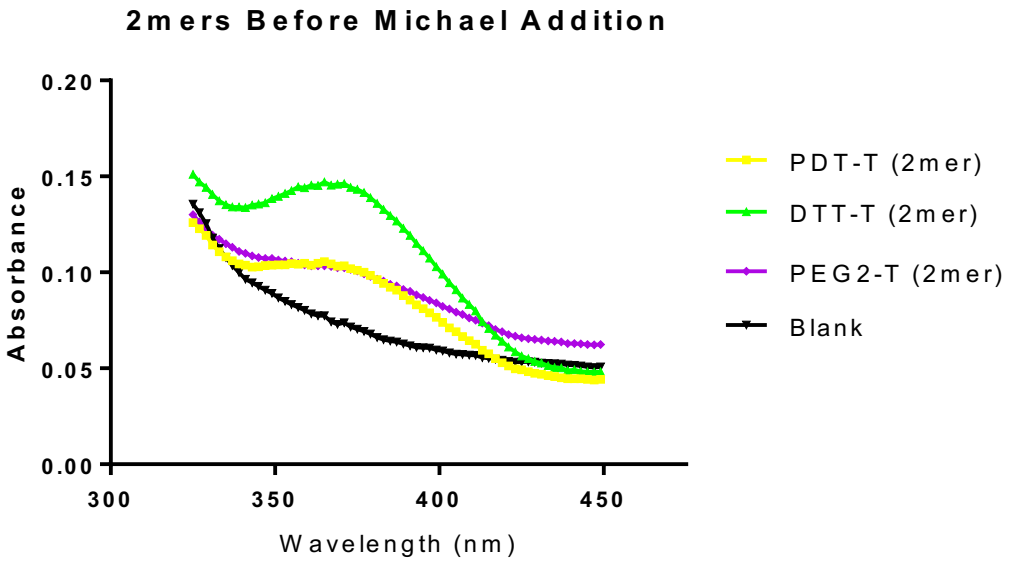

NMR Stack for Library Members with (D/L)-Dithiothreitol Comonomer
"Fluorous allyl amine"
DTT (0.5 mer)
DTT-T (1 mer)
DTT (1.5 mer)
DTT (2.5 mer)
DTT-T (3 mer)
6
5
4
3
2
1
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
f1 (ppm)

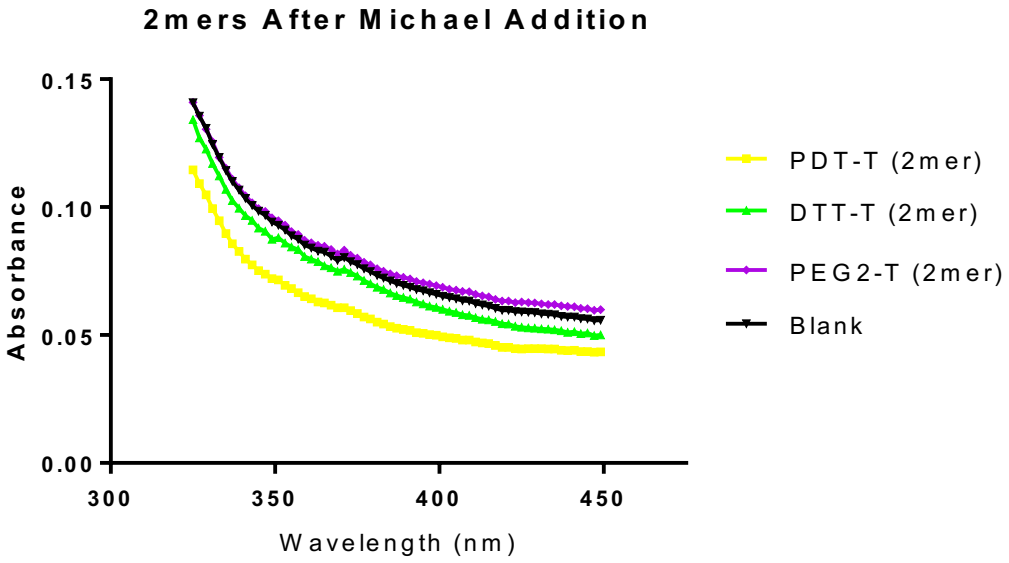

NMR Stack for Library Members with Propanedithiol Comonomer
"Fluorous allyl amine"
PDT (0.5 mer)
PDT-T (1 tmer)
PDT (1.5 mer)
PDT (2.5 mer)
PDT-T (3 mer)
6
5
4
3
2
1
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
f1 (ppm)

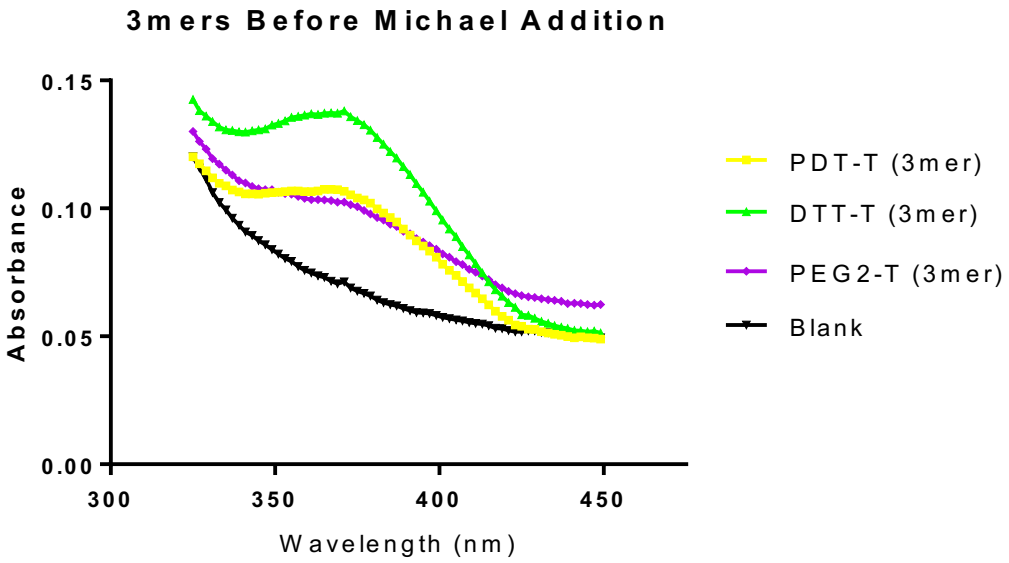
DTDP assay

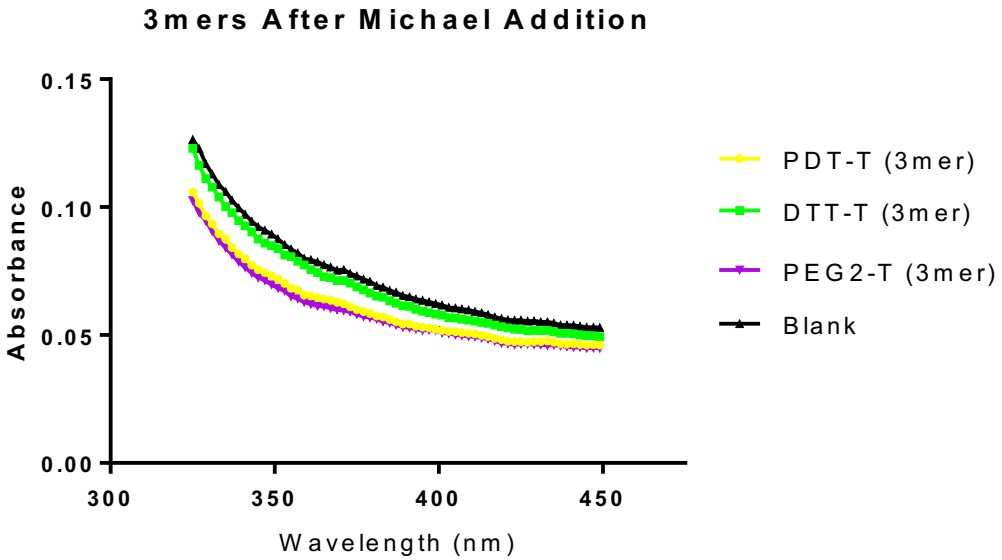

1mers Before Michael Addition
Absorbance
0.5
0.4
0.3
0.2
0.1
0.0
300
350
400
450
Wavelength (nm)
PDT-T (1mer)
DTT-T (1mer)
PEG2-T (1mer)
Blank


1mers After Michael Addition
Absorbance
0.20
0.15
0.10
0.05
0.00
300
350
400
450
Wavelength (nm)
PDT-T (1mer)
DTT-T (1mer)
PEG2-T (1mer)
Blank

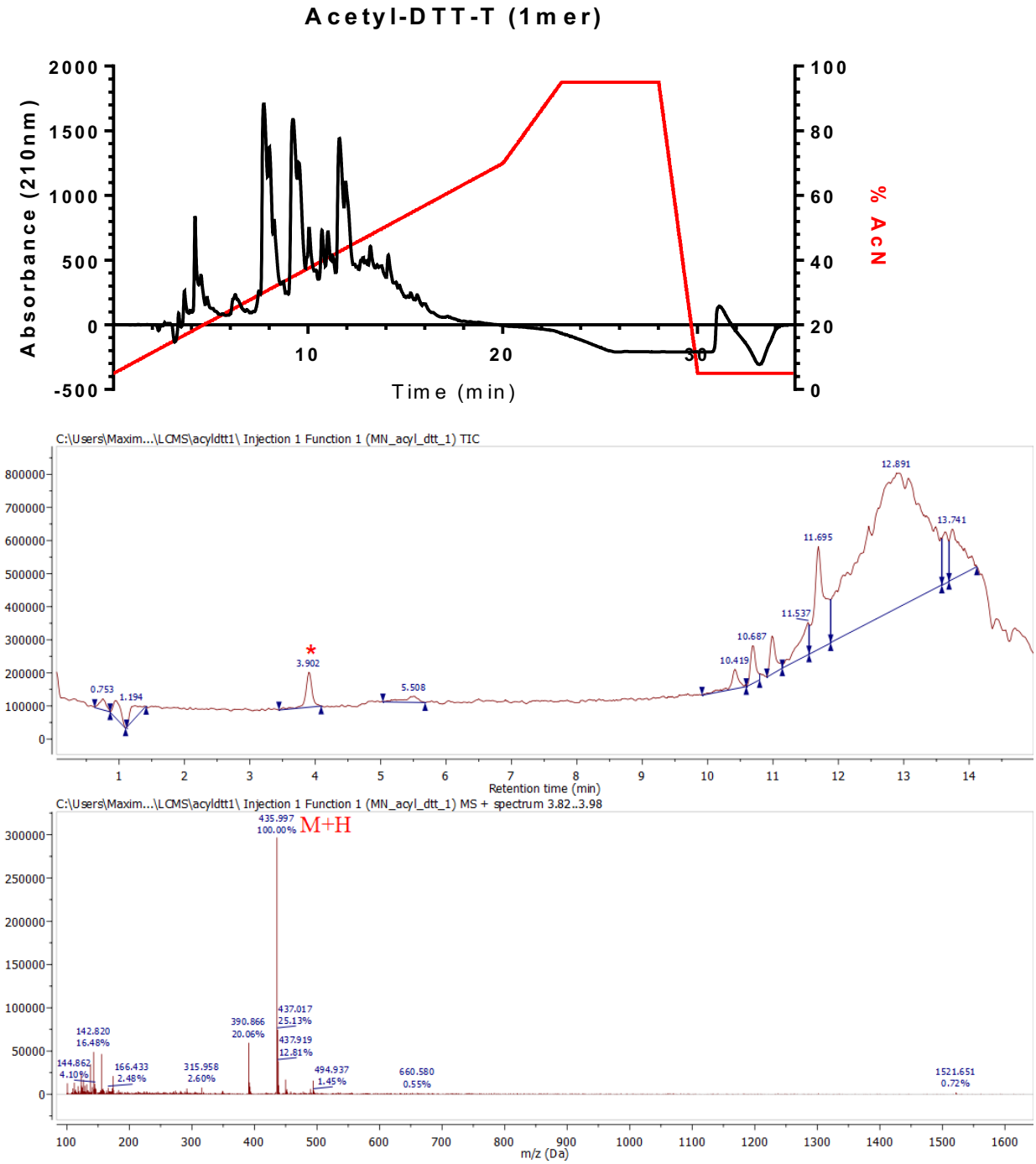

2mers Before Michael Addition
Absorbance
0.20
0.15
0.10
0.05
0.00
300
350
400
450
Wavelength (nm)
PDT-T (2mer)
DTT-T (2mer)
PEG2-T (2mer)
Blank

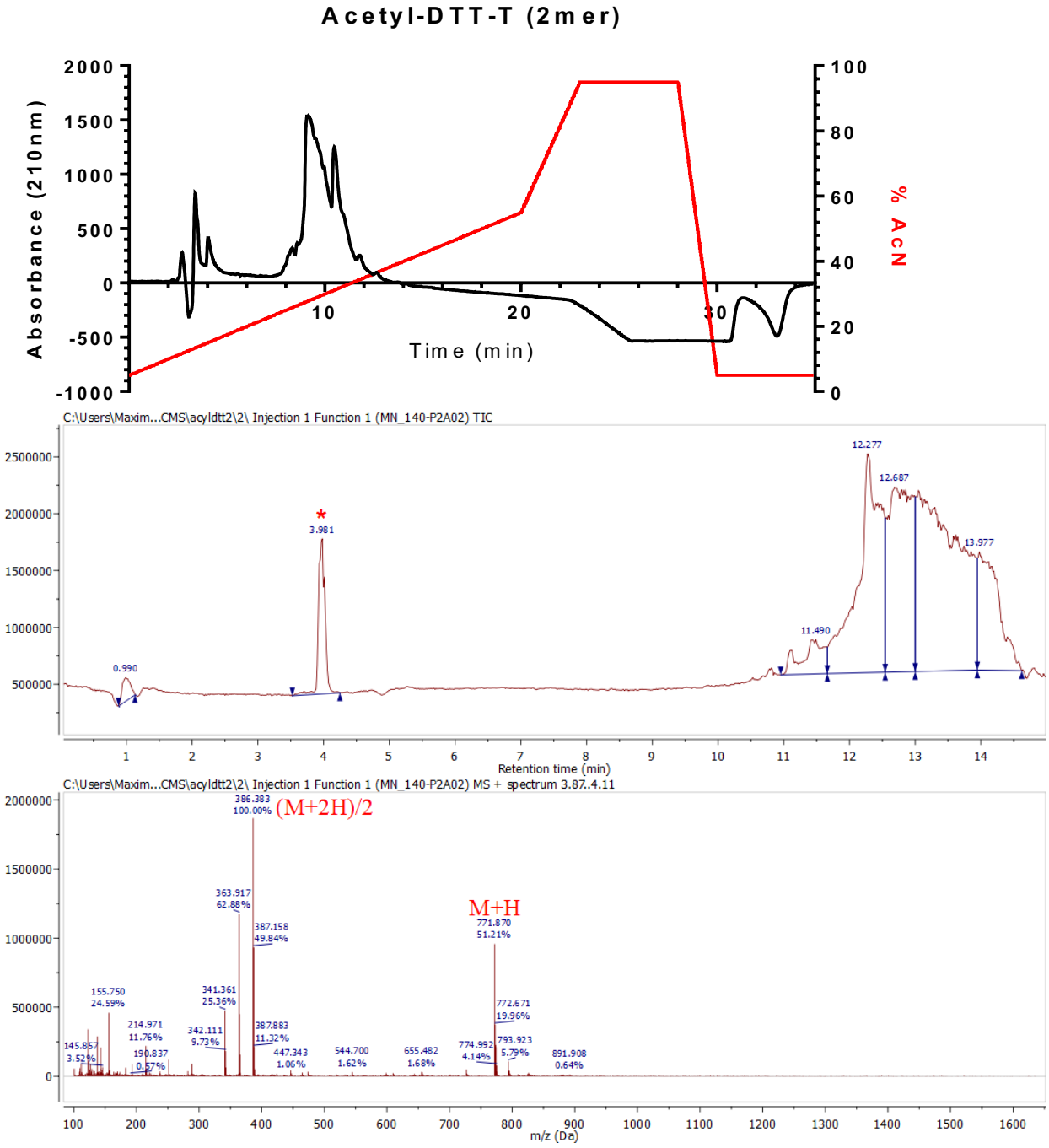

2mers After Michael Addition
Absorbance
0.15
0.10
0.05
0.00
300
350
400
450
Wavelength (nm)
PDT-T (2mer)
DTT-T (2mer)
PEG2-T (2mer)
Blank

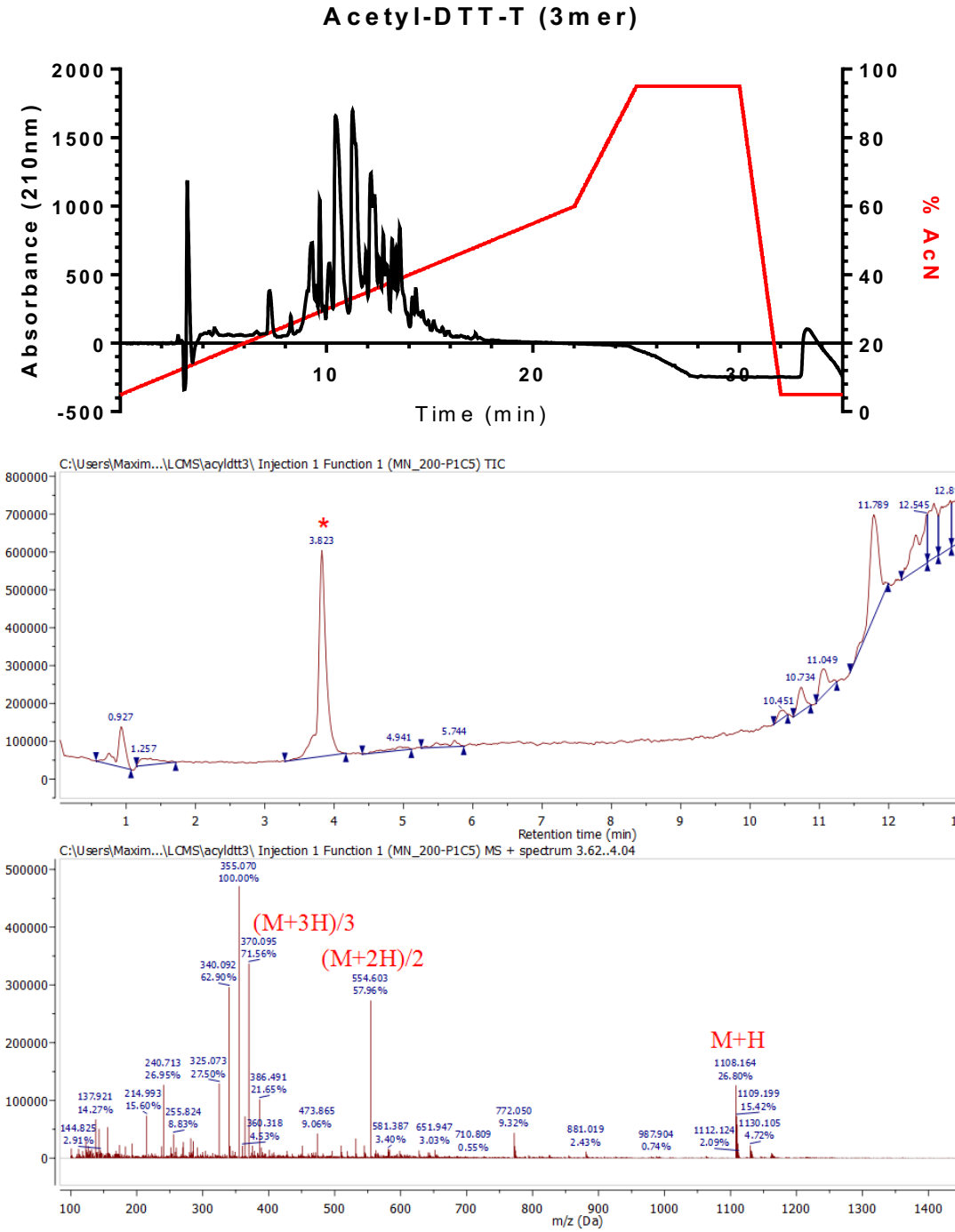

3mers Before Michael Addition
Absorbance
0.15
0.10
0.05
0.00
300
350
400
450
Wavelength (nm)
PDT-T (3mer)
DTT-T (3mer)
PEG2-T (3mer)
Blank

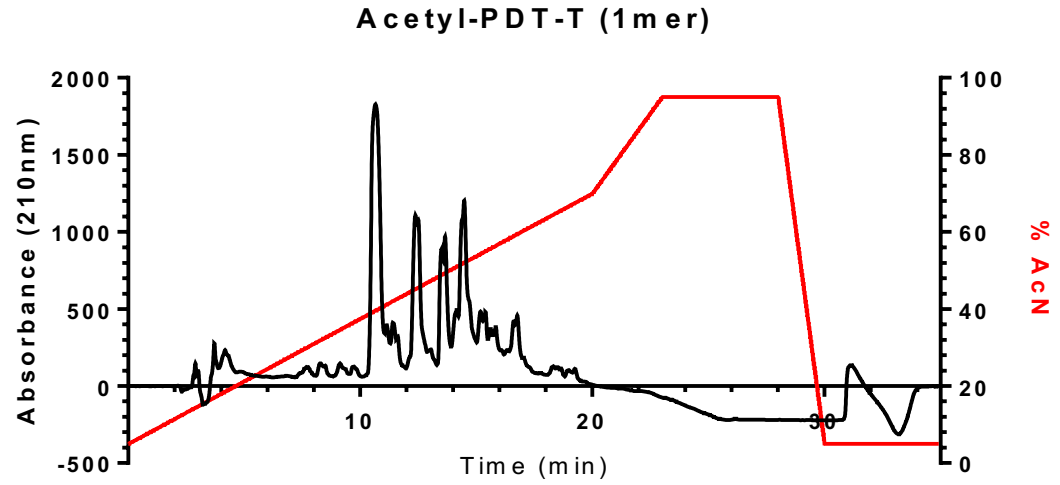

3mers After Michael Addition
Absorbance
0.15
0.10
0.05
0.00
300
350
400
450
Wavelength (nm)
PDT-T (3mer)
DTT-T (3mer)
PEG2-T (3mer)
Blank

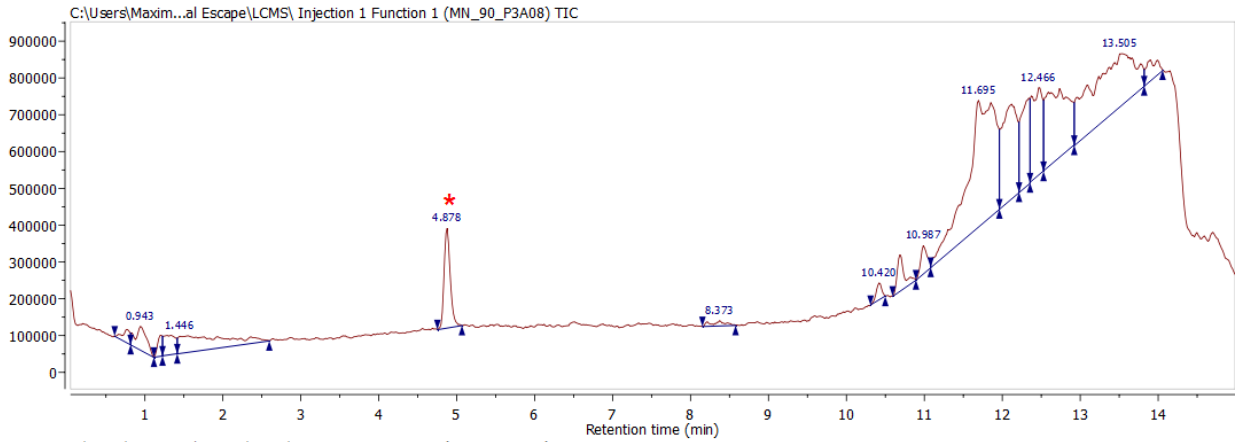
HPLC Traces and Corresponding LC-MS Traces and Mass Spectra

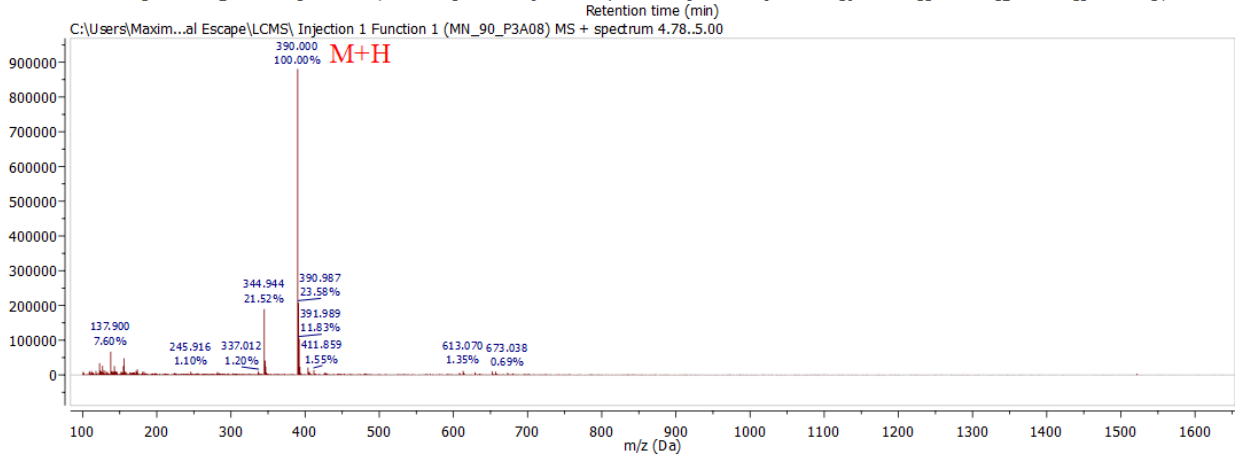


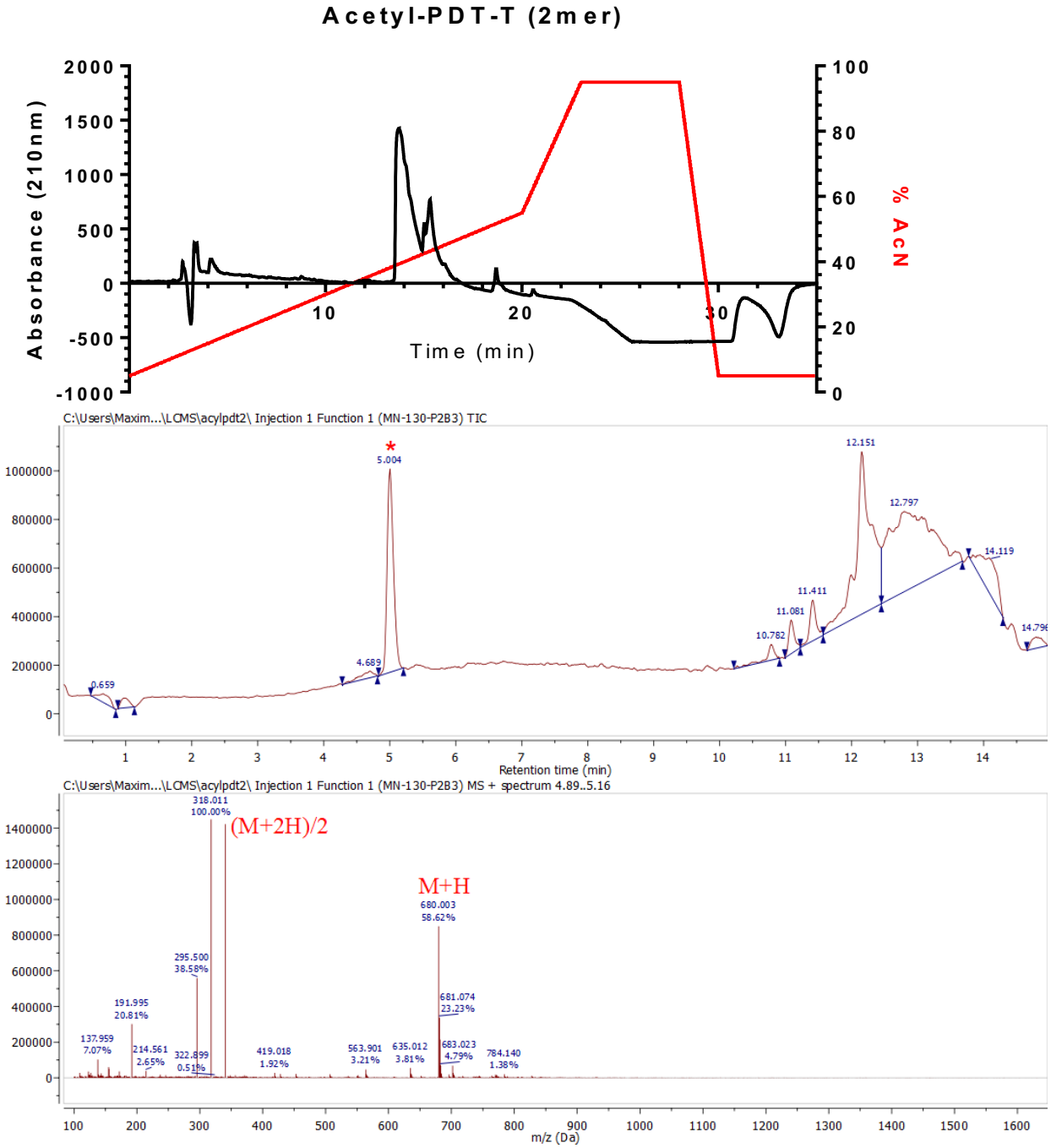

Acetyl-DTT-T (2mer)
Absorbance (210nm)
% AcN
Time (min)
C:\Users\Maxim...CMS\acyldtt2\2\ Injection 1 Function 1 (MN_140-P2A02) TIC
3.981
0.990
11.490
12.277
12.687
13.977
Retention time (min)
C:\Users\Maxim...CMS\acyldtt2\2\ Injection 1 Function 1 (MN_140-P2A02) MS + spectrum 3.87..4.11
386.383
100.00%
(M+2H)/2
363.917
62.88%
387.158
49.84%
M+H
771.870
51.21%
155.750
24.59%
341.361
25.36%
214.971
11.76%
342.111
9.73%
387.883
11.32%
772.671
19.96%
145.857
3.52%
190.837
0.57%
447.343
1.06%
544.700
1.62%
655.482
1.68%
774.992
4.14%
793.923
5.79%
891.908
0.64%
m/z (Da)

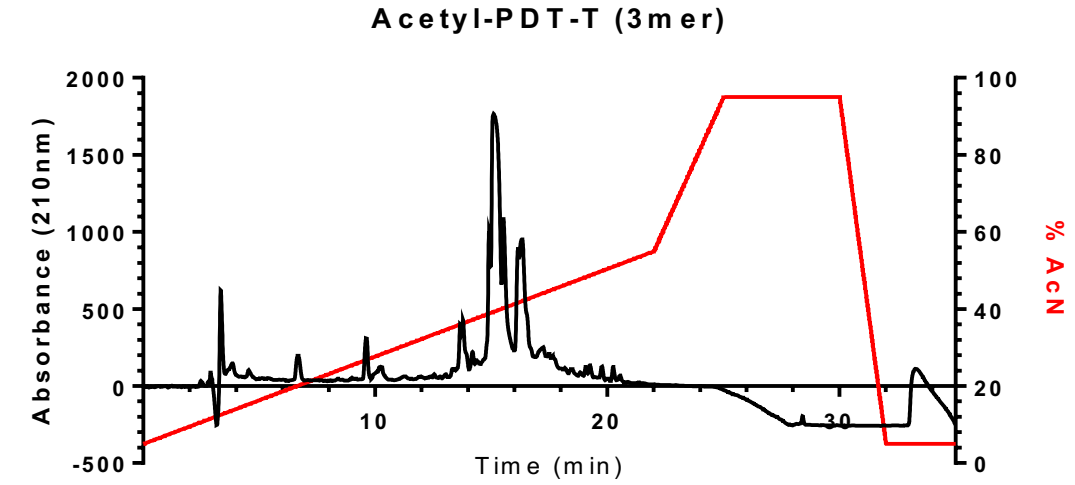

Acetyl-DTT-T (3mer)
Absorbance (210nm)
% AcN
Time (min)
C:\Users\Maxim...\LCMS\acyldtt3\ Injection 1 Function 1 (MN_200-P1C5) TIC
Retention time (min)
C:\Users\Maxim...\LCMS\acyldtt3\ Injection 1 Function 1 (MN_200-P1C5) MS + spectrum 3.62..4.04
(M+3H)/3
(M+2H)/2
M+H
m/z (Da)

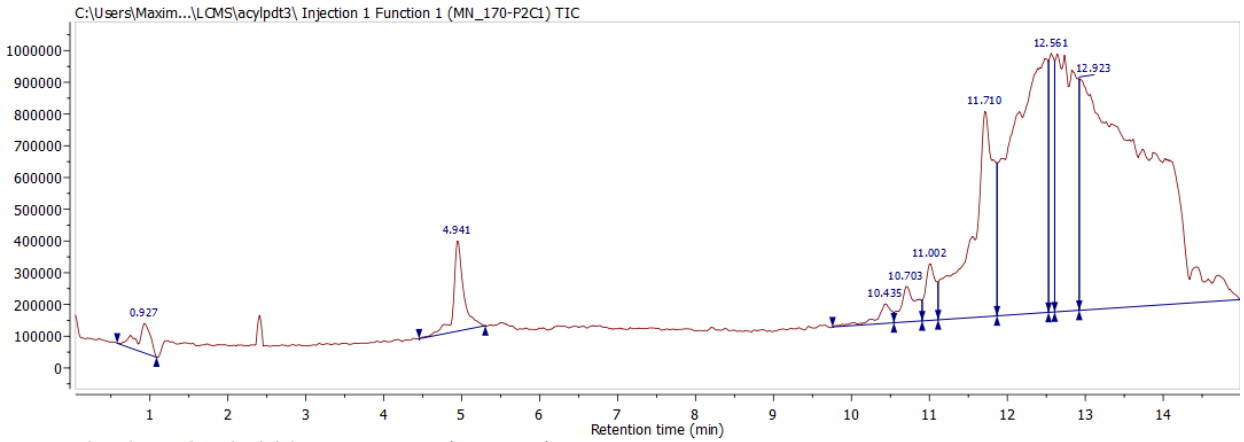

Acetyl-PDT-T (1mer)
Absorbance (210nm)
% AcN
Time (min)

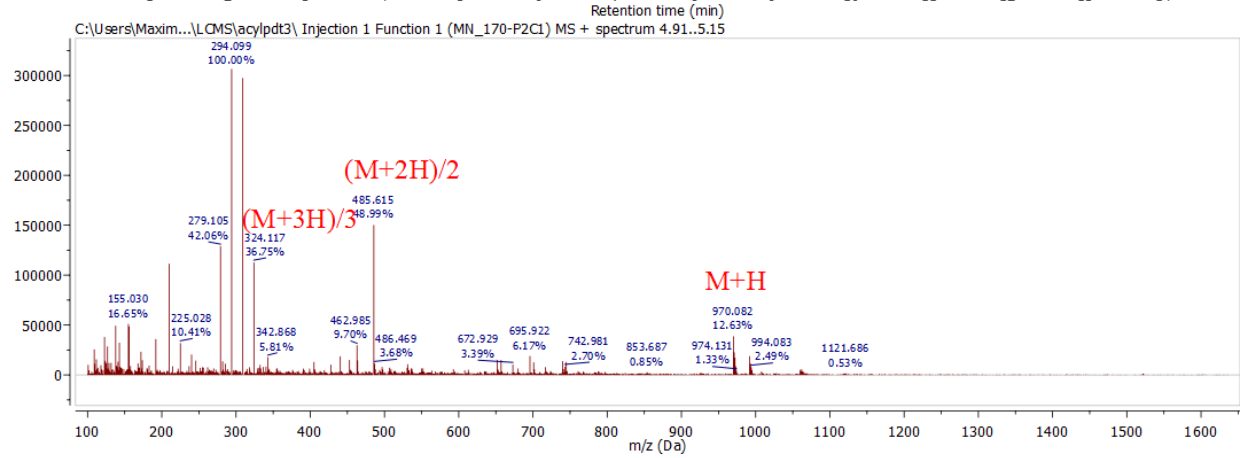

C:\Users\Maxim...al Escape\LCMS\ Injection 1 Function 1 (MN_90_P3A08) TIC
4.878
0.943
1.446
8.373
10.420
10.987
11.695
12.466
13.505
Retention time (min)

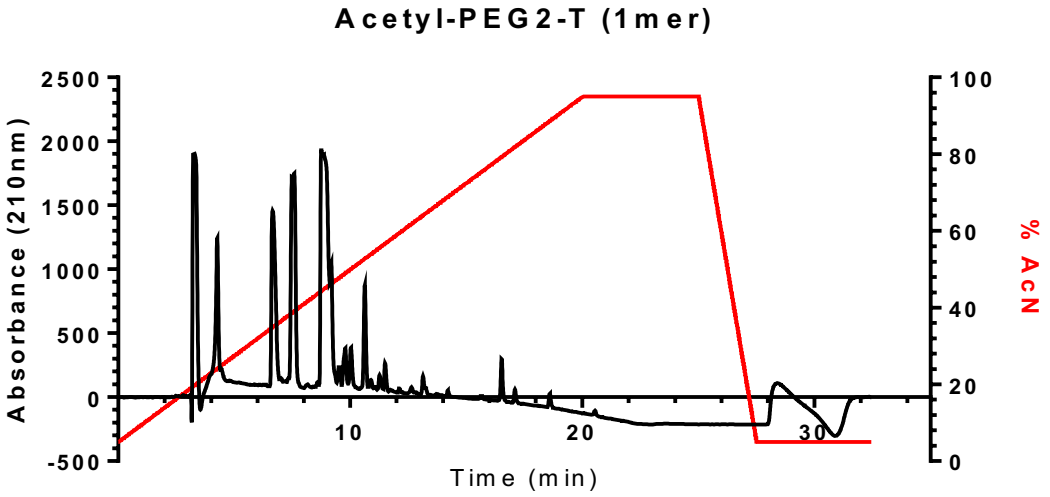

C:\Users\Maxim...al Escape\LCMS\ Injection 1 Function 1 (MN_90_P3A08) MS + spectrum 4.78..5.00
390.000
100.00%
M+H
344.944
21.52%
390.987
23.58%
391.989
11.83%
411.859
1.55%
137.900
7.60%
245.916
1.10%
337.012
1.20%
613.070
1.35%
673.038
0.69%
m/z (Da)

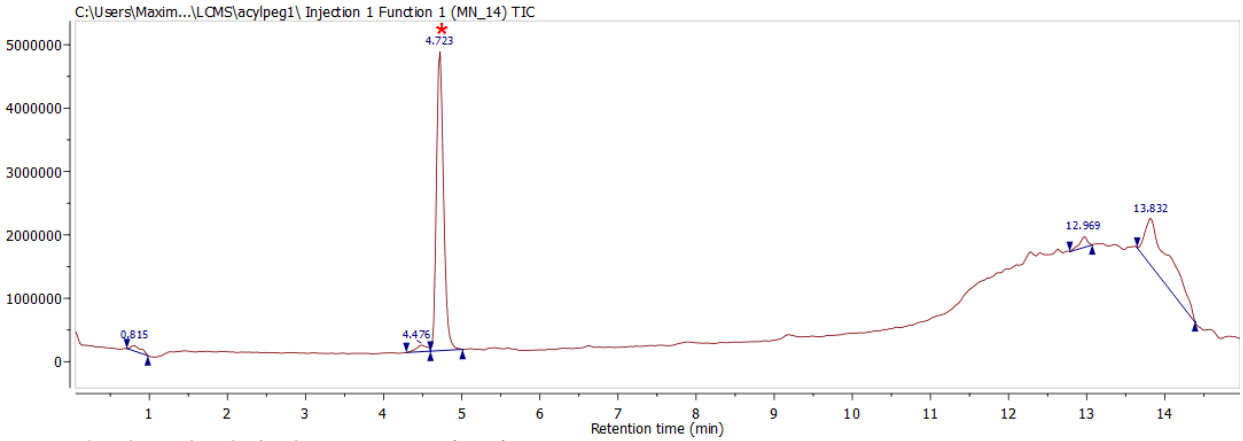

Acetyl-PDT-T (2mer)
Absorbance (210nm)
% AcN
Time (min)
C:\Users\Maxim...\LCMS\acylpdt2\ Injection 1 Function 1 (MN-130-P2B3) TIC
Retention time (min)
C:\Users\Maxim...\LCMS\acylpdt2\ Injection 1 Function 1 (MN-130-P2B3) MS + spectrum 4.89..5.16
(M+2H)/2
M+H
m/z (Da)

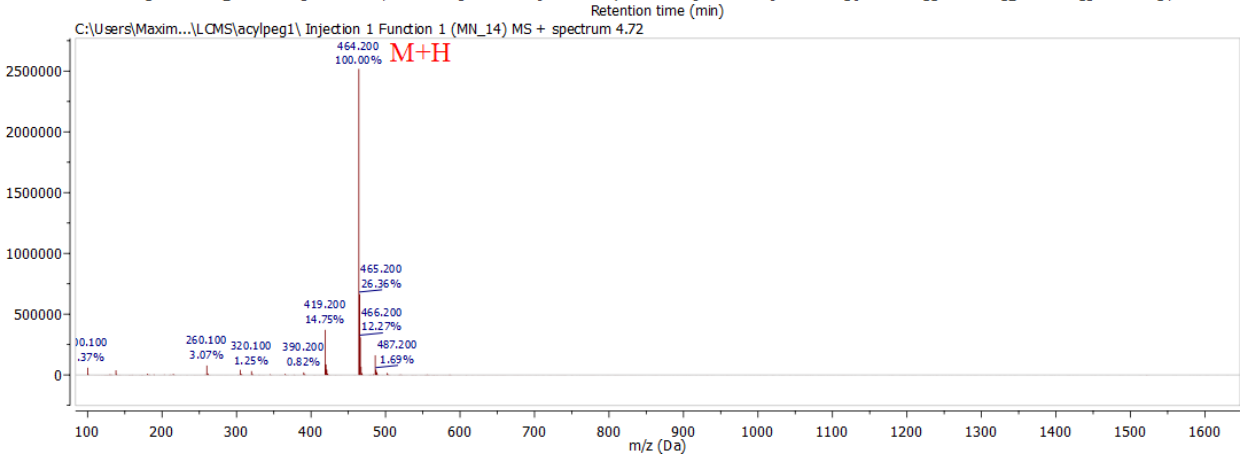

Acetyl-PDT-T (3mer)
Absorbance (210nm)
% AcN
Time (min)

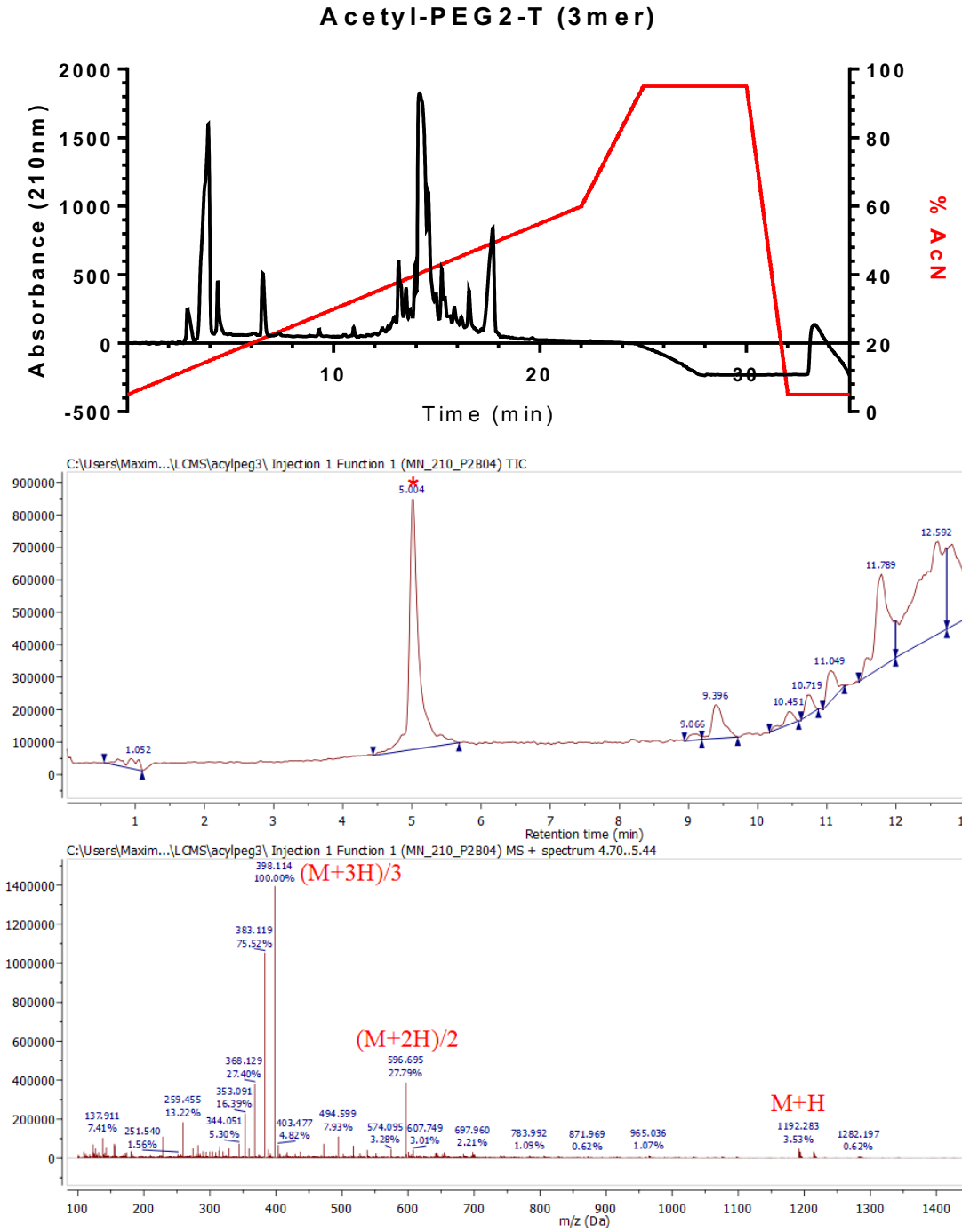

C:\Users\Maxim...\LCMS\acylpdt3\ Injection 1 Function 1 (MN_170-P2C1) TIC
0.927
4.941
10.435
10.703
11.002
11.710
12.561
12.923
Retention time (min)

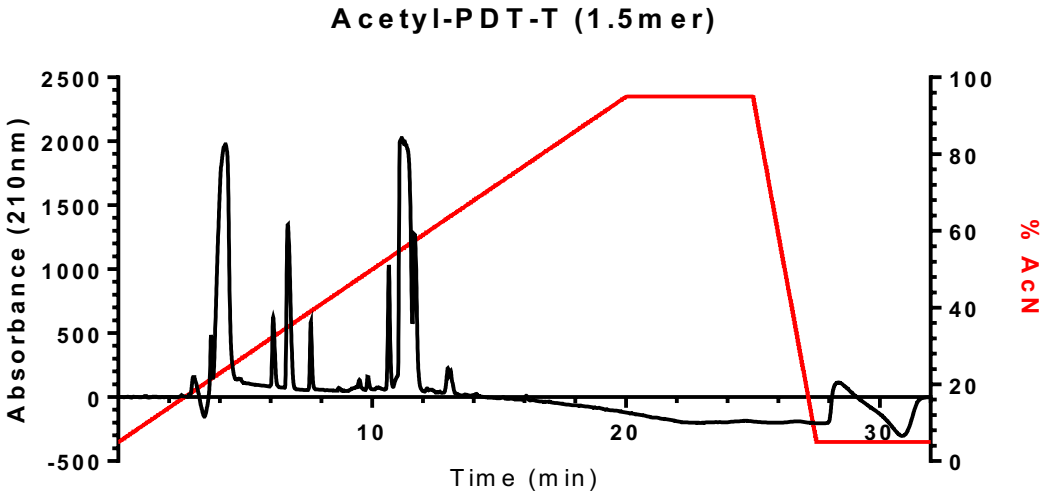

C:\Users\Maxim...\LCMS\acylpdt3\ Injection 1 Function 1 (MN_170-P2C1) MS + spectrum 4.91..5.15
294.099
100.00%
(M+2H)/2
485.615
48.99%
(M+3H)/3
279.105
42.06%
324.117
36.75%
M+H
970.082
12.63%
155.030
16.65%
225.028
10.41%
342.868
5.81%
462.985
9.70%
486.469
3.68%
672.929
3.39%
695.922
6.17%
742.981
2.70%
853.687
0.85%
974.131
1.33%
994.083
2.49%
1121.686
0.53%
m/z (Da)

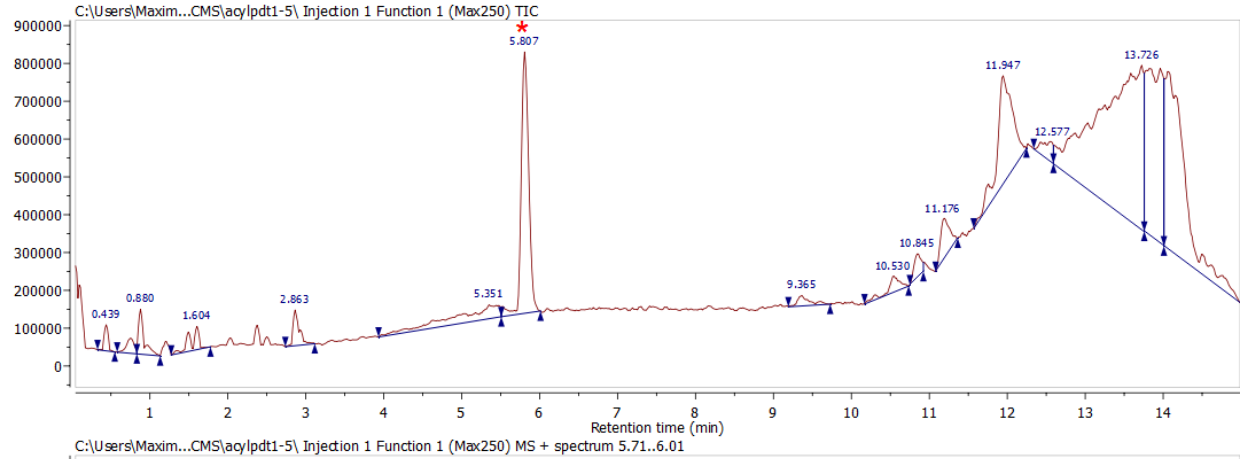

Acetyl-PEG2-T (1mer)
Absorbance (210nm)
% AcN
Time (min)

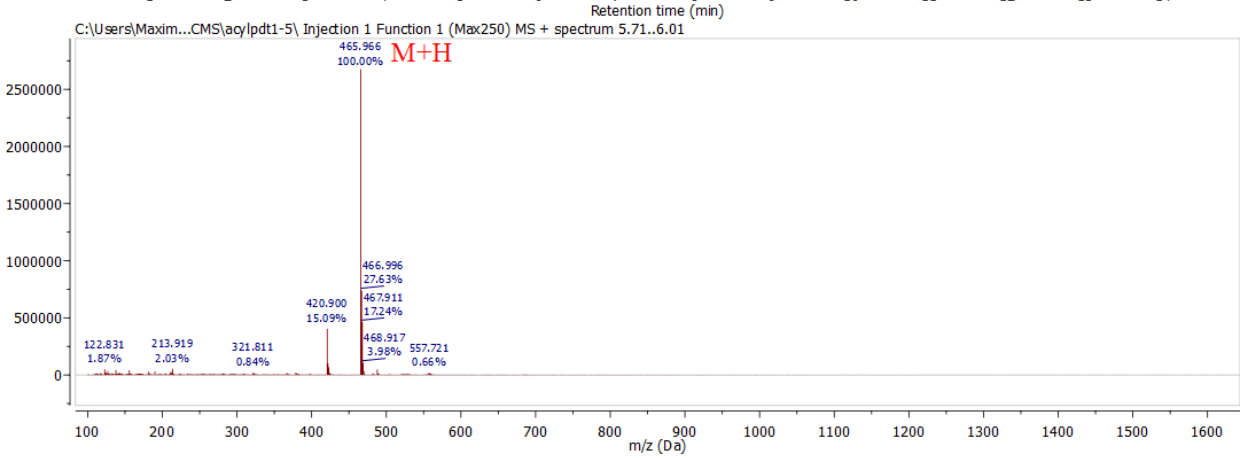

C:\Users\Maxim...\LCMS\acylpeg1\ Injection 1 Function 1 (MN_14) TIC
4.723
0.815
4.476
12.969
13.832
Retention time (min)

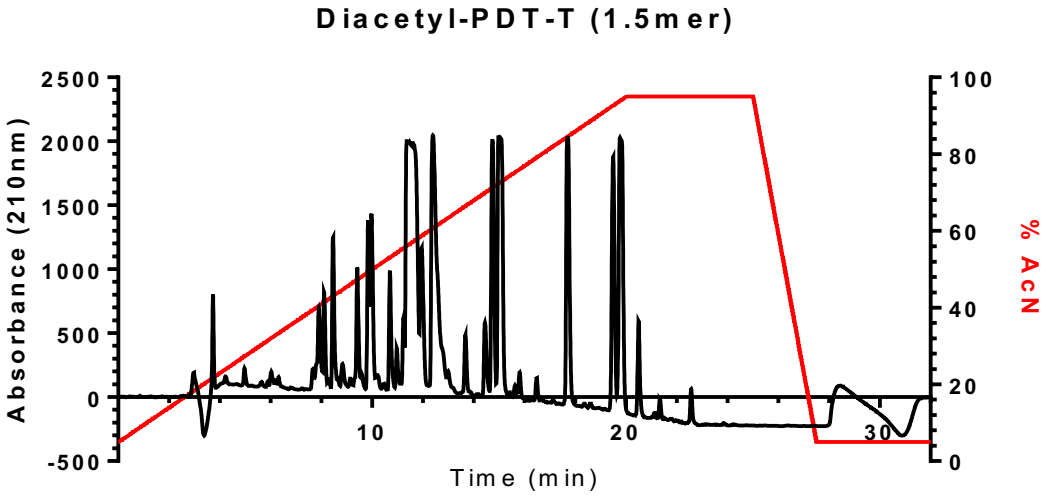

C:\Users\Maxim...\LCMS\acylpeg1\ Injection 1 Function 1 (MN_14) MS + spectrum 4.72
464.200
100.00%
M+H
465.200
26.36%
466.200
12.27%
419.200
14.75%
487.200
1.69%
260.100
3.07%
320.100
1.25%
390.200
0.82%
m/z (Da)

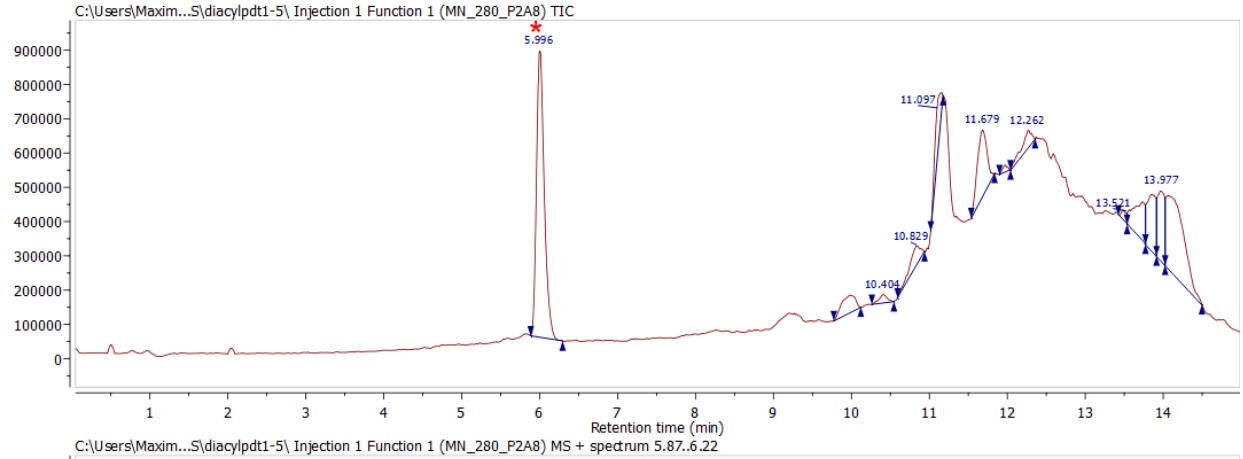

Acetyl-PEG2-T (3mer)
Absorbance (210nm)
% AcN
Time (min)
C:\Users\Maxim...\LCMS\acylpeg3\ Injection 1 Function 1 (MN_210_P2B04) TIC
Retention time (min)
C:\Users\Maxim...\LCMS\acylpeg3\ Injection 1 Function 1 (MN_210_P2B04) MS + spectrum 4.70..5.44
(M+3H)/3
(M+2H)/2
M+H
m/z (Da)

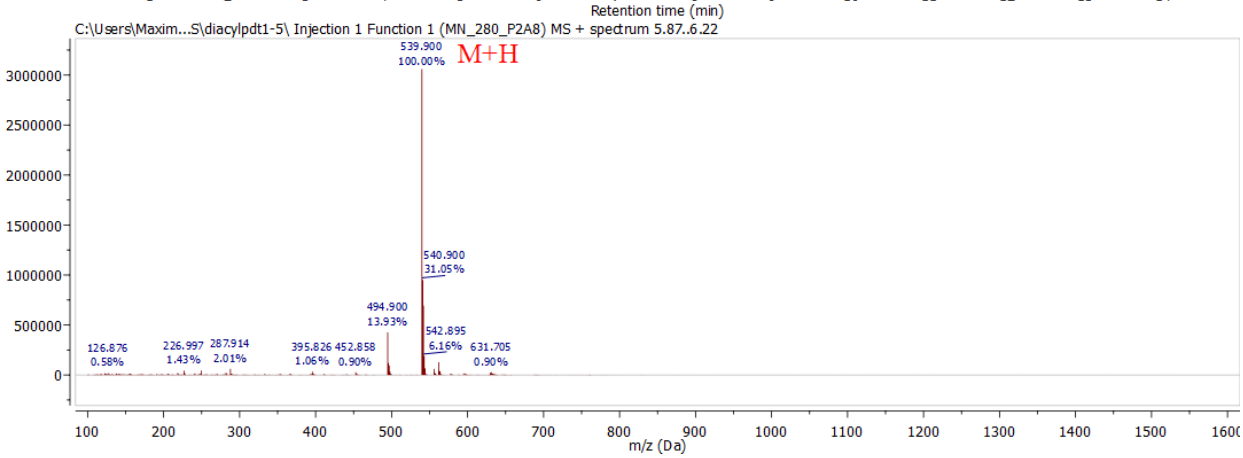

Acetyl-PDT-T (1.5mer)
Absorbance (210nm)
% AcN
Time (min)

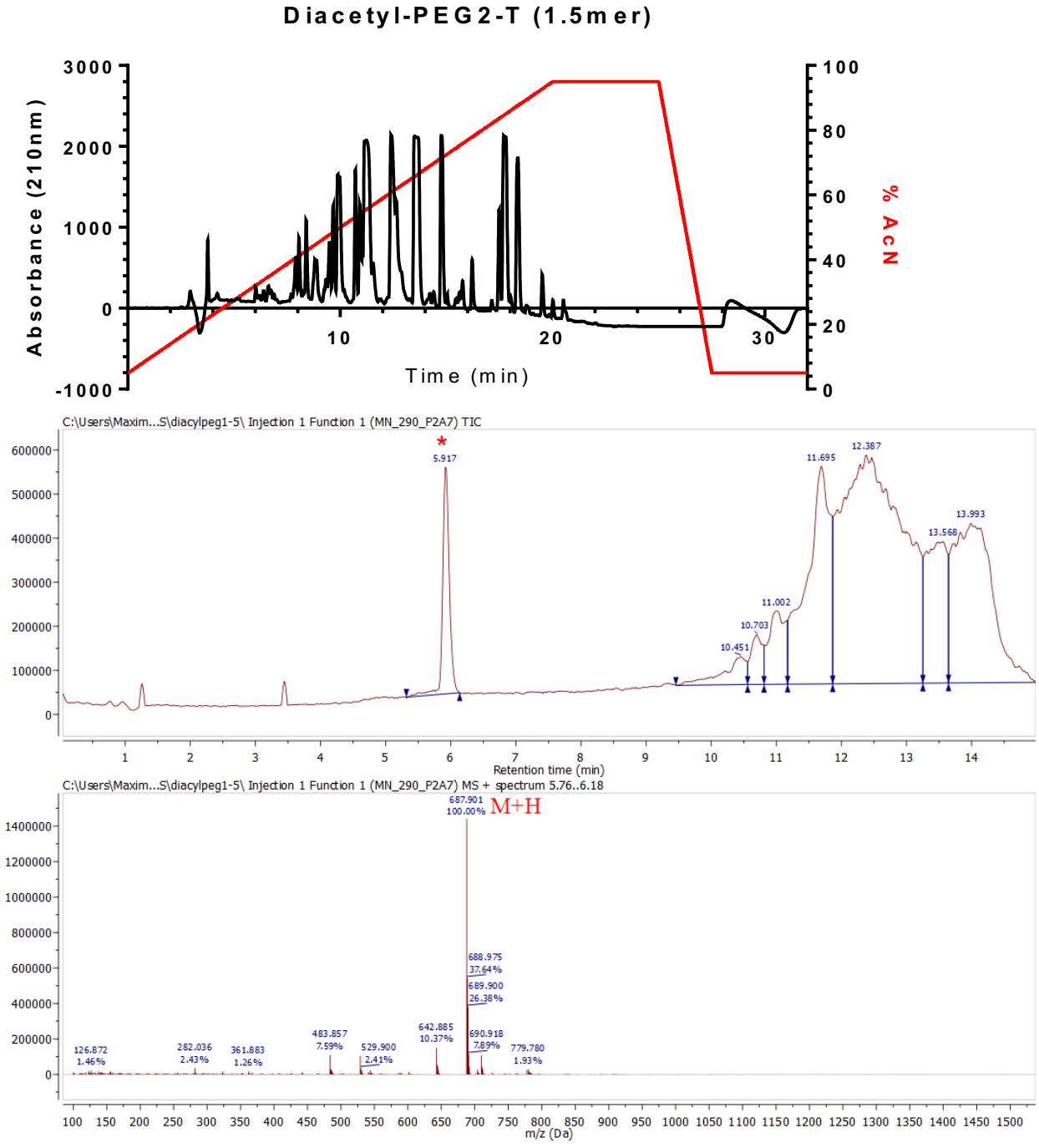

C:\Users\Maxim...CMS\acylpdt1-5\ Injection 1 Function 1 (Max250) TIC
5.807
11.947
13.726
12.577
11.176
10.845
10.530
9.365
5.351
2.863
1.604
0.880
0.439
Retention time (min)


C:\Users\Maxim...CMS\acylpdt1-5\ Injection 1 Function 1 (Max250) MS + spectrum 5.71..6.01
465.966
100.00%
M+H
466.996
27.63%
467.911
17.24%
468.917
3.98%
420.900
15.09%
122.831
1.87%
213.919
2.03%
321.811
0.84%
557.721
0.66%
m/z (Da)


Diacetyl-PDT-T (1.5mer)
Absorbance (210nm)
% AcN
Time (min)

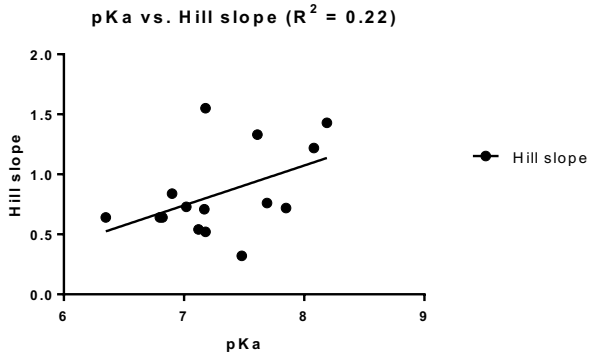

C:\Users\Maxim...S\diacylpdt1-5\ Injection 1 Function 1 (MN_280_P2A8) TIC
5.996
11.097
11.679
12.262
13.977
13.571
10.829
10.404
Retention time (min)

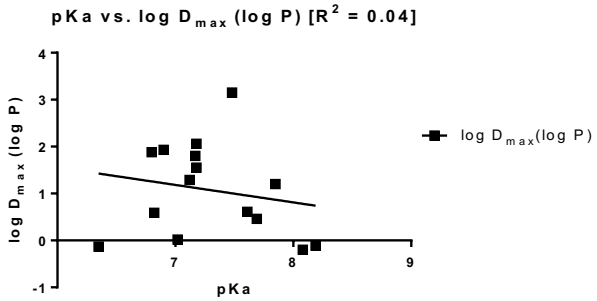

C:\Users\Maxim...S\diacylpdt1-5\ Injection 1 Function 1 (MN_280_P2A8) MS + spectrum 5.87..6.22
539.900
100.00%
M+H
540.900
31.05%
494.900
13.93%
542.895
6.16%
126.876
0.58%
226.997
1.43%
287.914
2.01%
395.826
1.06%
452.858
0.90%
631.705
0.90%
m/z (Da)

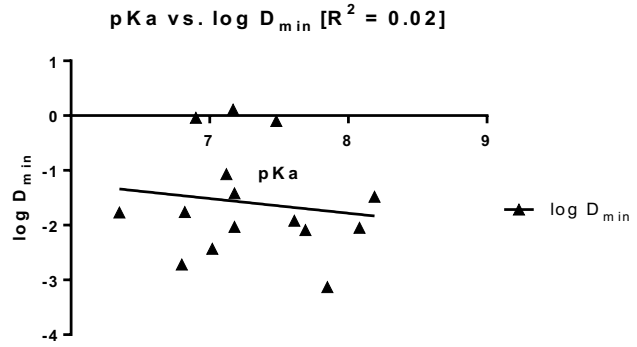

Diacetyl-PEG2-T (1.5mer)
Absorbance (210nm)
% AcN
Time (min)
C:\Users\Maxim...S\diacylpeg1-5\ Injection 1 Function 1 (MN_290_P2A7) TIC
5.917
10.451
10.703
11.002
11.695
12.387
13.568
13.993
Retention time (min)
C:\Users\Maxim...S\diacylpeg1-5\ Injection 1 Function 1 (MN_290_P2A7) MS + spectrum 5.76..6.18
687.901
100.00%
M+H
688.975
37.64%
689.900
26.38%
690.918
7.89%
126.872
1.46%
282.036
2.43%
361.883
1.26%
483.857
7.59%
529.900
2.41%
642.885
10.37%
779.780
1.93%
m/z (Da)

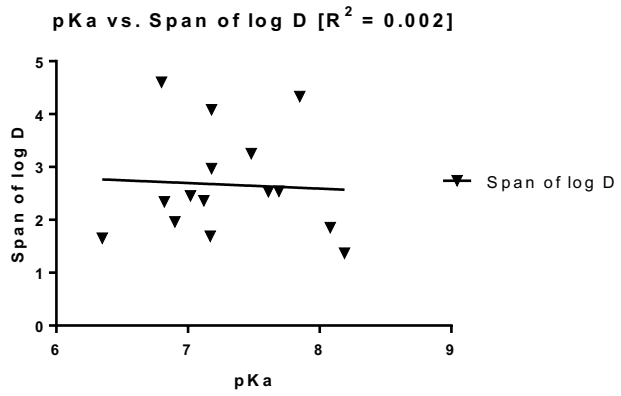
2-D Regression Plots between Parameters

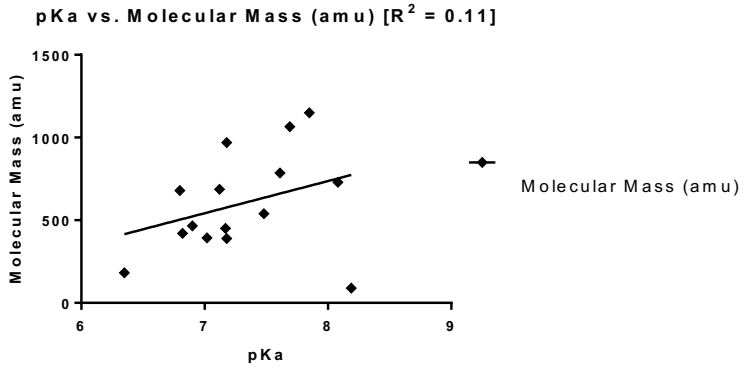
pK$_a$



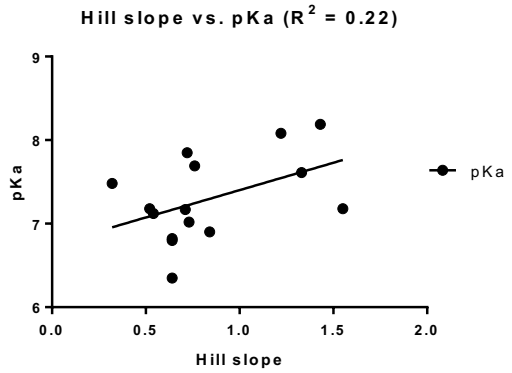


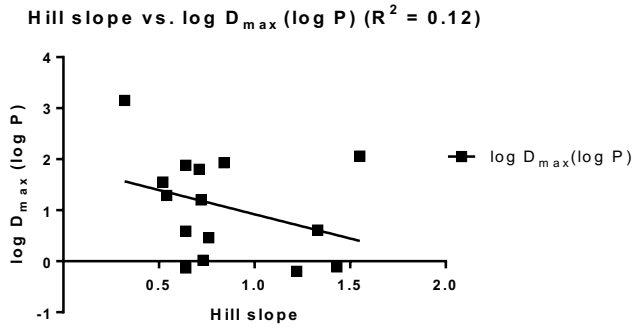


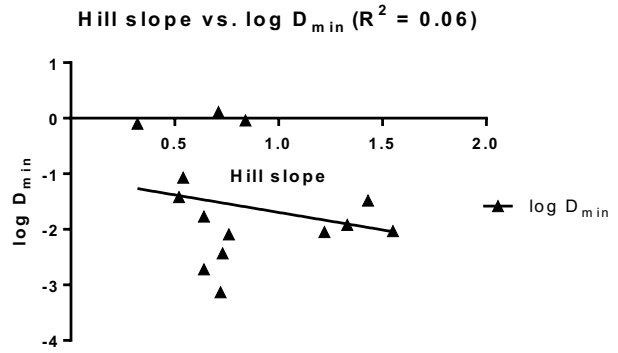

pKa vs. Span of log D [R2 = 0.002]
Span of log D
pKa
Span of log D
0
1
2
3
4
5
6
7
8
9

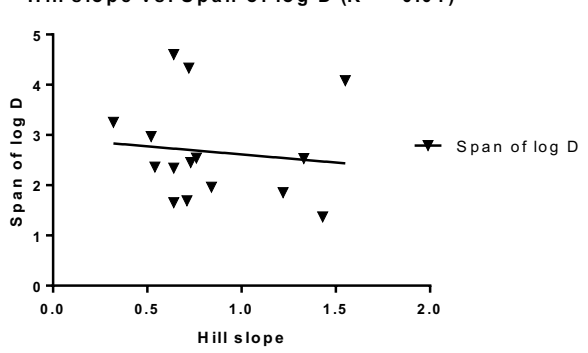

pKa vs. Molecular Mass (amu) [R2 = 0.11]
Molecular Mass (amu)
pKa
Molecular Mass (amu)
0
500
1000
1500
6
7
8
9

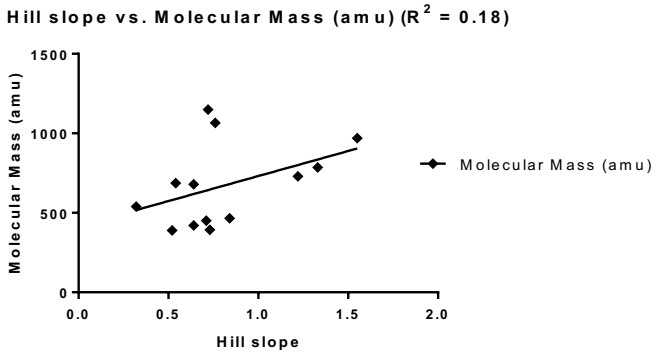
Hill Slope


Hill slope vs. pKa (R^2 = 0.22)
pKa
Hill slope
9
8
7
6
0.0
0.5
1.0
1.5
2.0

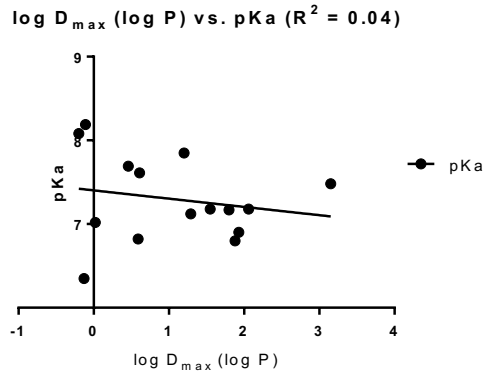

Hill slope vs. log D_{max} (log P) (R^2 = 0.12)
log D_{max} (log P)
log D_{max}(log P)
Hill slope
4
3
2
1
0
-1
0.5
1.0
1.5
2.0

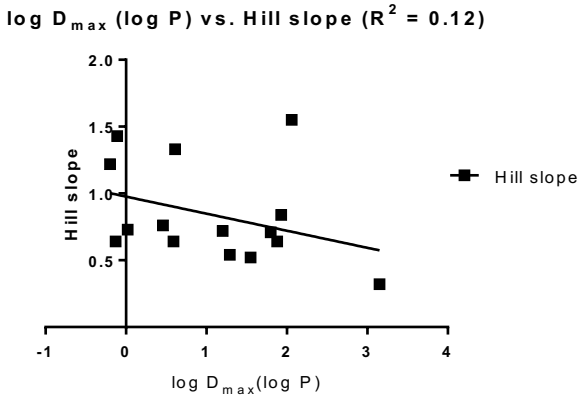

Hill slope vs. log D_{min} (R^2 = 0.06)
log D_{min}
log D_{min}
Hill slope
1
0
-1
-2
-3
-4
0.5
1.0
1.5
2.0

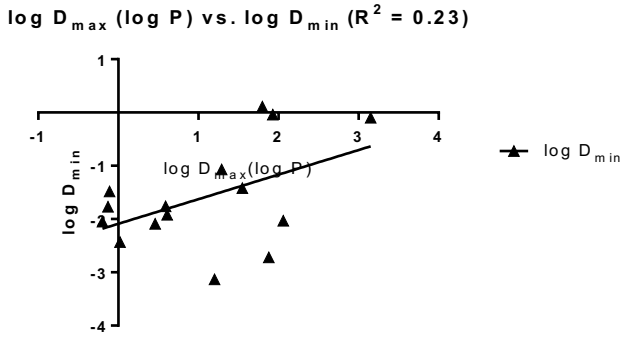

Hill slope vs. Span of log D (R^2 = 0.01)
Span of log D
Span of log D
Hill slope
5
4
3
2
1
0
0.0
0.5
1.0
1.5
2.0


Hill slope vs. Molecular Mass (amu) (R^2 = 0.18)
Molecular Mass (amu)
1500
1000
500
0
0.0
0.5
1.0
1.5
2.0
Hill slope
Molecular Mass (amu)

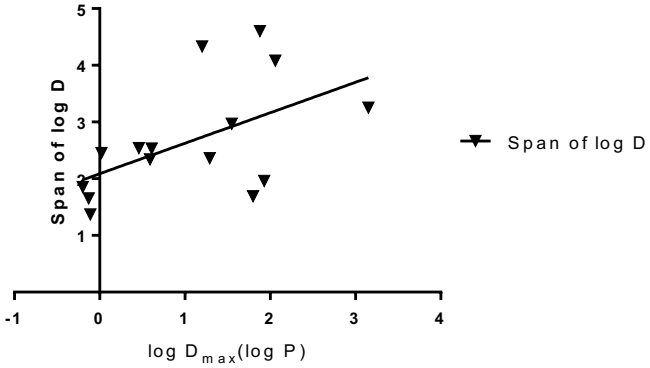
Log D_{max} (log P)



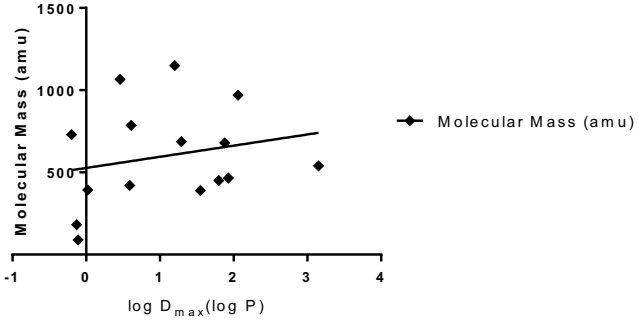




log D_{max} (log P) vs. Span of log D (R^2 = 0.28)

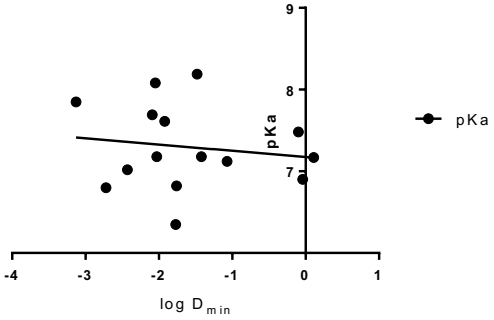


log D_{max} (log P) vs. Molecular Mass (amu) (R^2 = 0.05)

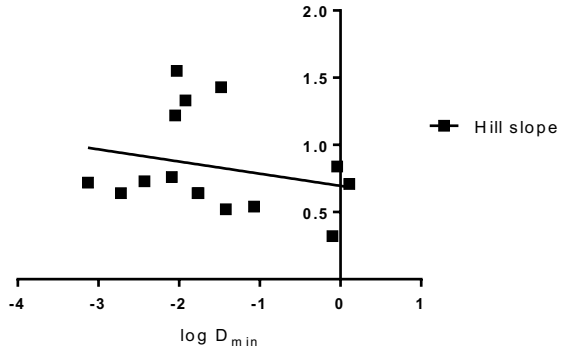


Log D_{min}

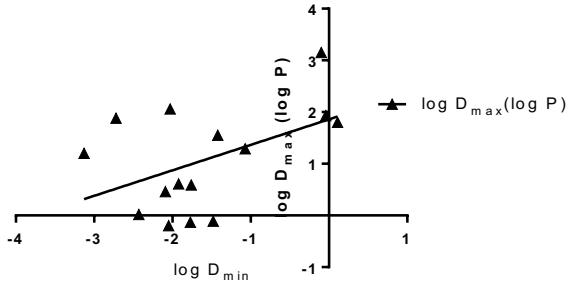
log D_{min} vs. pKa (R^2 = 0.02)



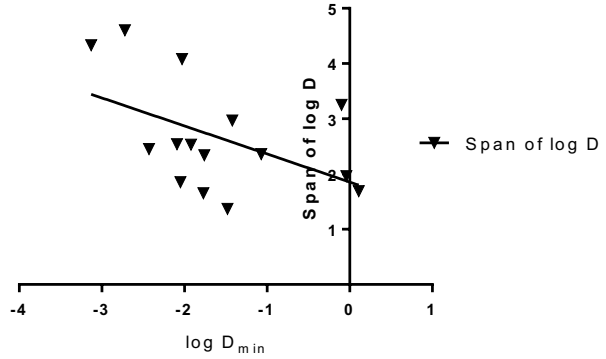
log D_{min} vs. Hill slope (R^2 = 0.06)



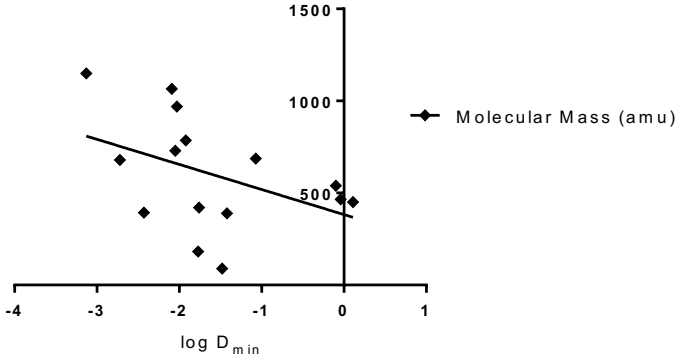
log D_{min} vs. log D_{max} (log P) (R^2 = 0.23)



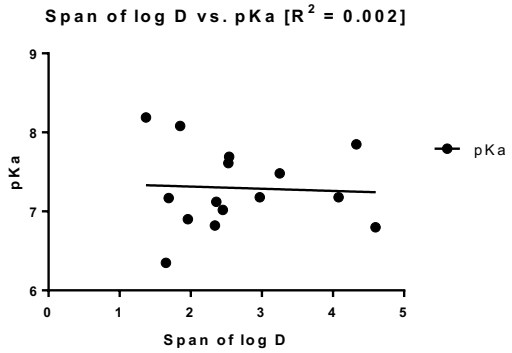
log D_{min} vs. Span of log D (R^2 = 0.24)

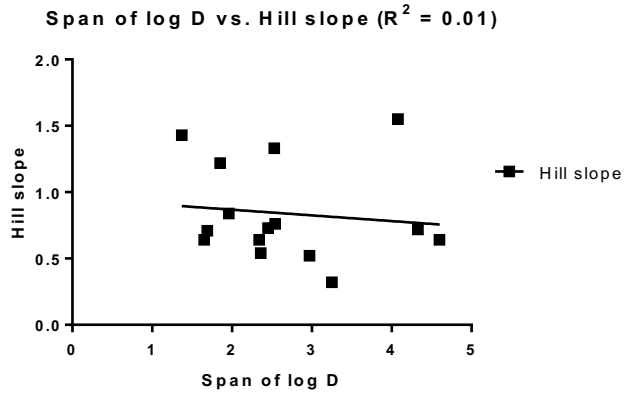


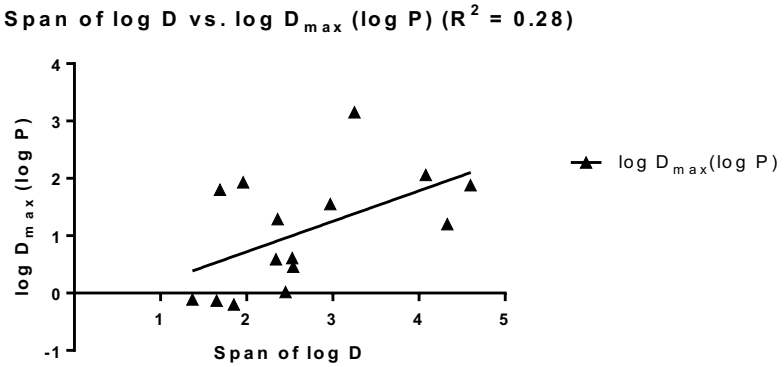
log D_{min} vs. Molecular Mass (amu) [R^2 = 0.18]



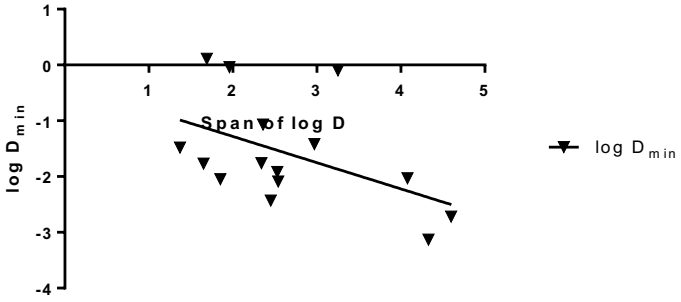
Span of log D


Span of log D vs. pKa [R2 = 0.002]
pKa
Span of log D
9
8
7
6
0
1
2
3
4
5
pKa

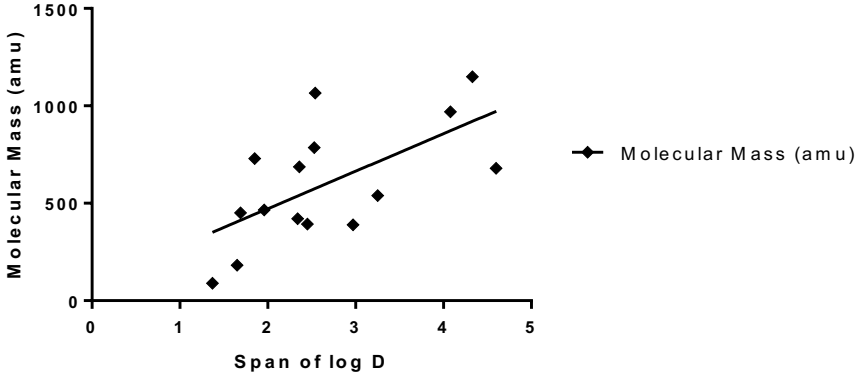

Span of log D vs. Hill slope (R2 = 0.01)
Hill slope
Span of log D
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
Hill slope


Span of log D vs. log Dmax (log P) (R2 = 0.28)
log Dmax (log P)
Span of log D
4
3
2
1
0
-1
1
2
3
4
5
log Dmax(log P)

Span of log D vs. log D_{min} (R^2 = 0.24)



Span of log D vs. Molecular Mass (amu) [R^2 = 0.39]



Comparing Models with AICc

The R^2 statistic is not appropriate for use when doing nonlinear regression. Instead, I will look at the Akaike information criteria (AIC)[56,57]. Specifically, we will look at AICc to correct for dealing with a relatively small sample size. The calculation for the difference between AIC for 2 models is as follows,

$$\Delta AIC_{\text{Model B}-\text{Model A}} = N \cdot \ln\left(\frac{\text{Sum}-\text{of}-\text{squares}_{\text{Model B}}}{\text{Sum}-\text{of}-\text{squares}_{\text{Model A}}}\right) + 2 \cdot \Delta k_{\text{Model B}-\text{Model A}} +$$

where N is the sample size, and k is the number of parameters in the model.

Roughly speaking, the value of this number estimates the relative quality of fit between models. It only provides a means of comparing the quality between models and does not give an absolute quality of the model. In other words, it can not indicate whether or not a model is the true underlying model that generated the data; rather, it serves as a way to pick the best model from a pool of potential model candidates. A more negative AICc value indicates a relatively better model. Grounded in principles from information theory, the AICc assesses goodness-of-fit while also comparing model complexity. With respect to the complexity, the criterion penalizes models for having too many parameters (overfitting).

With the difference in AICc between two models calculated, one can then estimate the relative probability with which one model is better. That calculation is as follows,

$$Relative\ Probability\ of\ the\ better\ Model = \frac{e^{0.5\cdot\Delta AICc}}{1+e^{0.5\cdot\Delta AICc}}$$

$$Relative\ Probability\ of\ the\ worse\ Model = 1 - Relative\ Probability\ of\ the\ better\ Model$$

The Boltzmann-like calculation for the probability reflects the idea that the minimization of AICc is analogous to maximizing the entropy in thermodynamics, highlighting again the information theory basis for this criterion.

Using PRISM's software, the different models for each oligoTEA in the library and their AICc are listed. For cases where 2 models had relatively similar AICc, a note was made on the probability ratio. In all the other cases, the best model has a probability of >99% than the 2nd best model. The best model(s) for each oligomer are highlighted in red.

Oligomer	AICc Value Model 1 (2 parameters)	AICc Value Model 2 (3 parameters)	AICc Value Model 3 (4 parameters)	Notes on comparing AICc values between models
Acetyl-DTT-T (1 mer)	**-102.8**	-100.1		Model 1 has a probability ratio of 3.82 over Model 2 (79% to 21%)

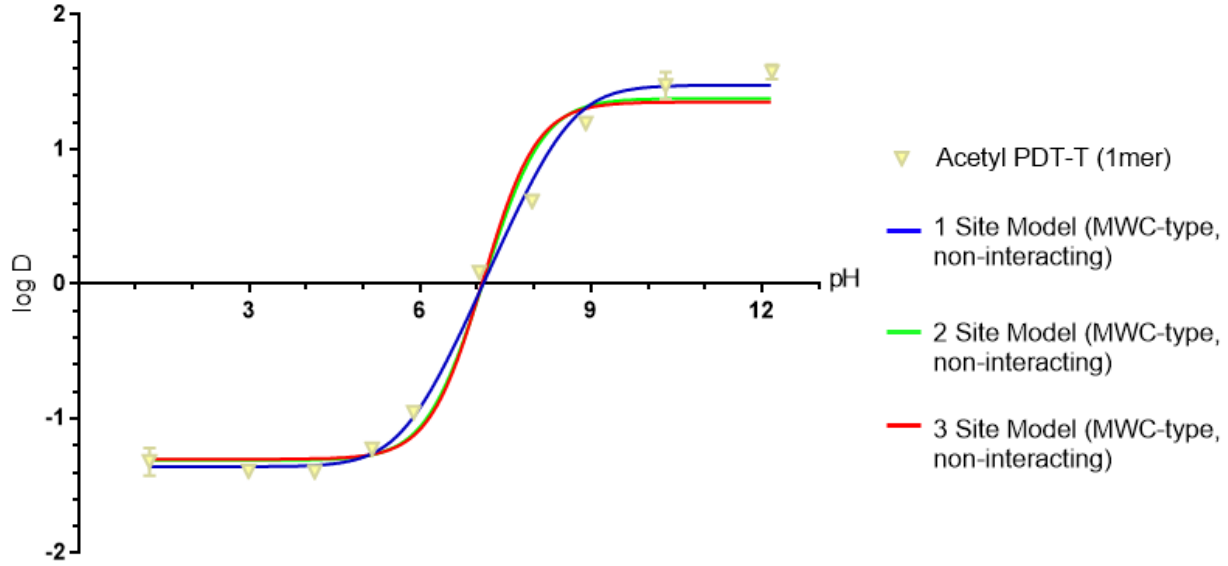
Acetyl-DTT-T (2 mer)	-175	-193.1	**-206.8**	
Acetyl-DTT-T (3 mer)	-31.27	**-34.41**	-31.4	Model 2 has a probability ratio of 4.49 over Model 3 (82% to 18%). Model 2 has a probability ratio of 4.81 over Model 1 (83% to 17%)
Acetyl-PDT-T (1 mer)	-0.798	**-83.82**		
Acetyl-PDT-T (2 mer)	11.46	**-110.8**	**-109.7**	Model 2 has a probability ratio of 1.72 over Model 3 (63% to 37%)
Acetyl-PDT-T (3 mer)	18.14	-67.86	**-118.6**	
Acetyl-PEG2-T (1 mer)	-77.68	**-155**		
Acetyl-PEG2-T (2 mer)	-31.39	-72.16	**-84.34**	
Acetyl-PEG2-T (3 mer)	-12.07	**-36.14**	**-34.52**	Model 2 has probability ratio of 2.25 over Model 3 (69% to 31%)

It is interesting to note that for only about half of the multi-mer oligomers Model 3 (which has the most parameters) is deemed the best model. Model 2 (the non-interacting MWC-like model) was objectively the best in several cases, edging out Model 3 for the DTT-T 3 mer, PDT-T 2 mer, and PEG2-T 3 mer. And Model 1 was even the best for the DTT-T 1mer. We see that the inclusion of cooperative effects is best for the 2 mer case in the DTT family, but when we increase length to the 3 mer, the non-cooperative model is seen to be a better model. This was also the case in the PEG2 family. Thus, the need to account for cooperative effects in a family of oligomers is not necessarily a linear function of its length.

Testing for Fitting Dependence on the Number of Protonation Sites

If Model has Excessive Number of Sites

One might suspect that it is possible to incorrectly fit the 1mer data with models that are designed for a higher number of protonation sites (ex. models with 2- and 3- protonation sites). Note this is not the same as overfitting because adding more protonation sites does not increase the number of parameters in the model; rather, it changes the weights and multiplicities of the states. To test this, we compared fits of experimental data for the Acetyl-PDT-T (1mer) with MWC-type, non-interacting models for 1-, 2-, and 3-protonation sites. The 1-protonation site model by definition is a non-interacting model, so the non-interacting models for the 2- and 3-protonate sites were used for a more direct comparison.

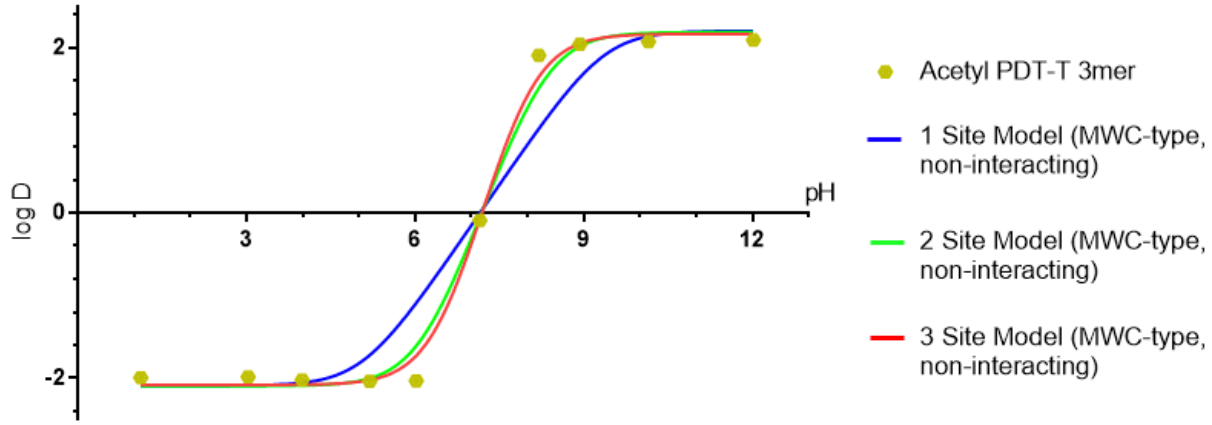


Number of Protonation Sites in a MWC-type, Non-Interacting Model	AICc
1-Protonation Site	-83.82
2-Protonation Sites	-62.12
3-Protonation Sites	-58.81

When one performs the fit, the best fitting MWC-type, non-interacting model for the experimental 1-mer data is actually the 1-protonation site model. This indicates that the model with the appropriate weights and multiplicities will best reflect the data, and that one cannot cheat the fitting by adding extra protonation sites into the model.

If Model Undercounts Number of Sites

If we look at the opposite scenario where the number of protonation sites in the model is undercounted relative to the number of sites in the actual compound, we still come to the same conclusion. Here we show that for an Acetyl PDT-T (3mer), the best fitting model is the one that has weights and multiplicities corresponding to 3 sites. The 1- and 2-protonation site models do not fit the data as cleanly.



Number of Protonation Sites in a MWC-type, Non-Interacting Model	AICc
1-Protonation Site	-21.68
2-Protonation Sites	-54.31
3-Protonation Sites	-67.86

Both of these examples illustrate that the Gibbs distribution correspond with accurate weights and multiplicities once you have defined the number of protonation states by properly enumerating the states.

Testing AICc on Overfitting of 1-mers

To test if this AICc could account for overfitting, I took the 1 mers and tried to apply Model 3 to them (i.e. a model with an interaction energy term). This physically doesn't make sense because a 1 mer only has 1 site and thus doesn't have any other site to physically interact with. But the addition of the parameter might still produce a good fit. Rather than having to rule out this model on physical intuition alone, I wanted to see if this criterion could do it for us objectively. Turns out it does it quite definitively. In all cases, the unphysical Model 3 that overfits the data is not deemed the best model by the AICc statistic.

	AICc Value			
Oligomer	**Model 1**	**Model 2**	**Model 3**	**Notes on AICc comparison between models**
Acetyl-DTT-T (1 mer)	**-102.8**	-100.1	-96.47	Model 1 has a probability ratio of 3.82 over Model 2 (79% to 21%). Model 1 has a probability ratio of 23.32 over Model 3 (96% to 4%). Model 3 also doesn't have a proper physical interpretation
Acetyl-PDT-T (1 mer)	-0.798	**-83.82**	-80.2	Model 2 has a probability ratio of 6.11 over Model 3 (86% to 14%). Model 3 also doesn't have a proper physical interpretation
Acetyl-PEG2-T (1 mer)	-77.68	**-155**	-152.4	Model 2 has probability ratio of 3.71 over Model 3 (79% to 21%). Model 3 also doesn't have a proper physical interpretation